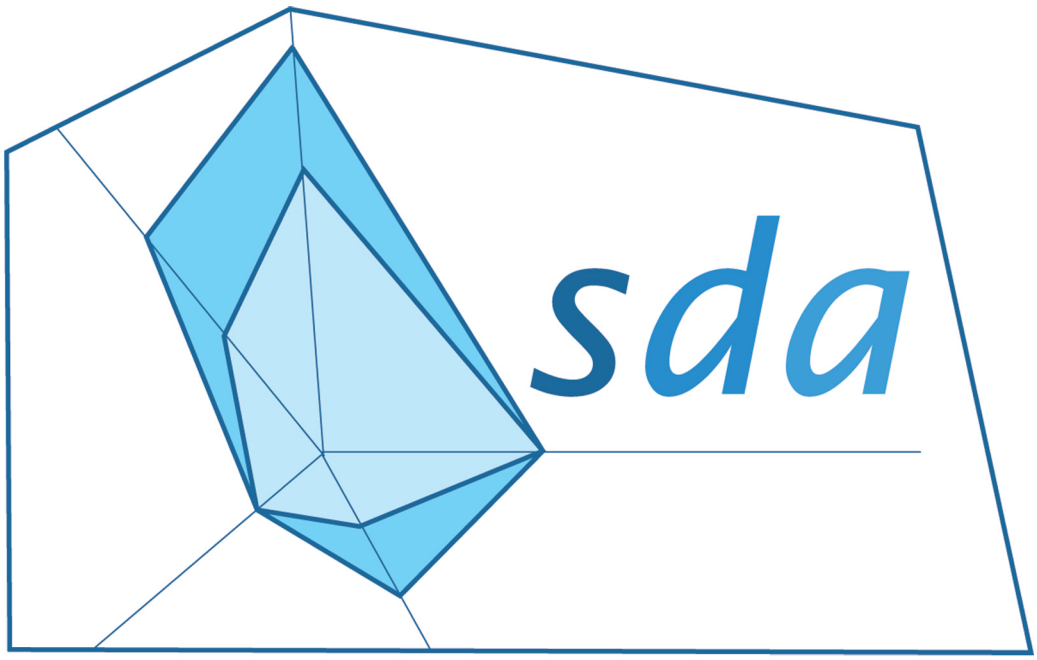
Javier Arroyo
Carlos Maté
Paula Brito
Monique Noirhomme-Fraiture (Eds.)

3rd Workshop in Symbolic Data Analysis

Madrid, 7-9 November 2012

Book of Abstracts



3rd Workshop in Symbolic Data Analysis

Universidad Complutense
de Madrid

J. Arroyo
C. Maté
P. Brito
M. Noirhomme-Fraiture (eds.)

November, 2012

Documento maquetado con TEXIS v.1.0+.

3rd Workshop in Symbolic Data Analysis

Book of Abstracts

November, 2012

ISBN 978-84-695-6575-9

Depósito legal: M-38333-2012

Preface

The organizers of the 3rd Workshop in Symbolic Data Analysis are pleased to present this book that gathers the abstracts of the works that will be presented during the Workshop. This workshop is the third regular meeting of researchers interested in Symbolic Data Analysis. The main aim of the event is to favor the meeting of people and the exchange of ideas from different fields - Mathematics, Statistics, Computer Science, Engineering, Economics, among others - that contribute to Symbolic Data Analysis.

We would like to thank everyone who contributed to make this workshop possible. The authors for their work, that makes Symbolic Data Analysis progress in its theoretical foundations and empirical applications; the IFCS and the EGC for sponsoring and disseminating the call for contributions; the Faculty of Computer Sciences of the Complutense University of Madrid and specially the Department of Software Engineering and Artificial Intelligence for their help and support; and especially the ICAI School of Engineering of Comillas Pontifical University of Madrid, through its Department of Industrial Organization and Institute for Research in Technology, for its help and support as well, and for making available the installations where this workshop is taking place.

Madrid, November 2012

Javier Arroyo
Universidad Complutense de Madrid, Spain

Carlos Maté
Universidad Pontificia Comillas, Spain

Paula Brito
Universidade do Porto, Portugal

Monique Noirhomme-Fraiture
Facultés Universitaires Notre-Dame de la Paix, Belgium

Contents

Session I

Clustering I

November 7, 11:30 - 13:00

Divisive Monothetic Clustering for Interval and Histogram-Valued Data

Paula Brito[1,⋆], **Marie Chavent**[2]

1. Faculdade de Economia & LIAAD-INESC TEC, Universidade do Porto, Portugal;
2. IMB et INRIA CQFD, Université de Bordeaux 2, Bordeaux, France
⋆ Contact author: mpbrito@fep.up.pt

Keywords: Divisive clustering, Histogram data, Interval data, Monothetic clustering.

We present a divisive top-down clustering method which is designed for interval and histogram-valued data (see, for instance, Diday (2000) or Brito (2011)). The method produces a hierarchy on a set of objects together with a monothetic characterization of each formed cluster. Interval-valued variables being a special case of histogram-valued variables, the method applies to data described by either kind of variables, or by variables of both types (see also Chavent (1998) and Chavent (2012)).

Let $\Omega = \{\omega_1, \ldots, \omega_n\}$ be the set under analysis. Histogram-valued variables Y_j are defined by a mapping $Y_j : \Omega \rightarrow B$ where B is the set of probability of frequency distribuitions on a finite set of intervals $\left\{I_{ij1}, ..., I_{ijk_{ij}}\right\}$ such that $Y_j(\omega_i) = \left(I_{ij1}, p_{ij1}; \ldots; I_{ijk_{ij}}, p_{ijk_{ij}}\right)$; $p_{ij\ell}$ is the probability or frequency associated to the sub-interval $I_{ij\ell} = \left[\underline{I}_{ij\ell}, \overline{I}_{ij\ell}\right[$ and $p_{ij1} + \ldots + p_{ijk_{ij}} = 1$.

Divisive clustering algorithms proceed top-down, starting with the set Ω to be clustered and performing a bi-partition of one cluster at each step. At step m a partition of Ω in m classes is present, one of which will be further split in two sub-classes; the class to be split and the split rule are chosen so as to produce a partition in $m+1$ classes that minimizes the internal class dispersion. The "quality" of a given partition $P_m = \left\{C_1^{(m)}, C_2^{(m)}, \ldots, C_m^{(m)}\right\}$ is measured by a criterion $Q(m)$, sum of the internal dipersion of each class: $Q(m) = \sum_{\alpha=1}^{m} I(C_\alpha) = \sum_{\alpha=1}^{m} \frac{1}{2n_\alpha} \sum_{\omega_i, \omega_h \in C_\alpha^{(m)}} D^2(\omega_i, \omega_h)$ with $D^2(\omega_i, \omega_h) = \sum_{j=1}^{p} d^2(x_{ij}, x_{hj})$ where d is a quadratic distance between distributions (note that both for interval and histogram-valued variables, $Y_j(\omega_i)$ may be represented by a distribution). For each class, internal dispersion is measured by the sum D^2 of the squared distances between all pairs of class members. We consider distances D^2 which are additive on the descriptive variables. At each step a class is selected to be split in two sub-classes, so as to minimize $Q(m+1)$ or, equivalently, maximize $Q(m) - Q(m+1)$ (note that Q decreases at each step).

Different distance measures may be considered to compare distribuitions. Let $Y_j(\omega_i) = H_{Y_j(\omega_i)} = ([\underline{I}_{ij1}, \overline{I}_{ij1}[, p_{ij1}; \ldots; [\underline{I}_{ijK_j}, \overline{I}_{ijK_j}], p_{ijK_j})$. We shall use the Mallows distance: $d_M^2(x_{ij}, x_{hj}) = \int_0^1 (q_{ij}(t) - q_{hj}(t))^2 dt$, where q_{ij} is the quantile function of the distribution $Y_j(\omega_i)$ or the Squared Euclidean distance: $d_E^2(x_{ij}, x_{hj}) = \sum_{\ell=1}^{K_j} (p_{ij\ell} - p_{hj\ell})^2$ (which imposes a same partition in sub-intervals for each observation of each variable Y_j).

The bi-partition to be made at each step is defined by a single variable, considering conditions of the type $R_{j\ell} := Y_j \leq \overline{I}_{j\ell}, \ell = 1, \ldots, K_j - 1, j = 1, \ldots, p$, which lead to a bi-partition of a class separating the elements that meet the given condition from the remainng ones. An element $\omega_i \in \Omega$

meets condition $R_{j\ell} = Y_j \leq \bar{I}_{j\ell}$ if and only if $\sum_{h=1}^{\ell} p_{ijh} \geq 0.5$. At each step, the class $C_\alpha^{(m)}$ and the condition $R_{j\ell}$ are selected such that the resulting partition P_{m+1}, in $m+1$ classes, minimizes $Q(m+1)$.

In the obtained clustering, each class is therefore represented by a conjunction of properties on the descriptive variables; the sequence of conditions met by the members of each class constitute necessary and sufficient conditions for class membership.

An example on social and crime data in the USA, where microdata recorded for towns has been aggregated by state using histogram-valued variables, illustrates the proposed method.

Acknowledgments

This work is partly funded by the ERDF European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP - 01-0124-FEDER-022701.

References

Bock, H.-H. and Diday, E. (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, Berlin-Heidelberg.

Brito, P. and Chavent, M. (2012). Divisive Monothetic Clustering for Interval and Histogram-Valued Data. In *Proc. ICPRAM 2012*, Vilamoura, Portugal.

Chavent, M. (1998). A monothetic clustering method. *Pattern Recognition Letters*, *19(11)*, 989–996.

Noirhomme-Fraiture, M. and Brito, P. (2011). Far beyond the classical data models: Symbolic data analysis. *Statistical Analysis and Data Mining*, *4(2)*, 157–170.

Clustering of Modal Valued Symbolic Data

Vladimir Batagelj[1], **Simona Korenjak Černe**[2], **Nataša Kejžar**[3]

1. University of Ljubljana, Faculty of Mathematics and Physics
2. University of Ljubljana, Faculty of Economics
3. University of Ljubljana, Faculty of Medicine
*Contact author: vladimir.batagelj@fmf.uni-lj.si

Keywords: Modal symbolic objects, Leaders method, Hierarchical clustering, Ward's method, Associations

Symbolic Data Analysis is based on special descriptions of data — symbolic objects (SO). Such descriptions preserve more detailed information about units and their clusters than the usual representations with mean values. A special kind of symbolic object is also a representation with frequency or probability distributions of modal values. This representation enables us to consider in the clustering process the variables of all measurement types at the same time.

In our previous talks a clustering criterion function for SOs was proposed such that the representative of each cluster is again composed of distributions of variables over the cluster. The corresponding leaders clustering method is based on this result. It is also shown that for the corresponding agglomerative hierarchical method a generalized Ward's formula holds. Both methods are compatible – they are solving the same clustering optimization problem.

The leaders method enables us to efficiently solve clustering problems with large number of units; while the agglomerative method can be applied alone on the smaller data set, or it could be applied on leaders, obtained with compatible nonhierarchical clustering method. Such a combination of two compatible methods enables us to decide upon the right number of clusters on the basis of the corresponding dendrogram.

In this talk we present an application of the proposed approach to *The European Social Survey data sets* (ESS, 2012). We also discuss the problem of identification of characteristic properties of the obtained clusters (Wikipedia, 2012) and analysis of associations among symbolic variables (Studer et al., 2011; Gabadinho et al., 2011).

References

ESS (2012). The European Social Survey. `http://ess.nsd.uib.no/`.

Gabadinho A., Studer M., Müller N.S., Ritschard G., Bürgin R. (2012). Trajectory miner: a toolbox for exploring and rendering sequence data. Package *TraMineR*. Version 1.8-2, June 4, 2012. `http://cran.r-project.org/web/packages/TraMineR/TraMineR.pdf`.

Studer M., Ritschard G., Gabadinho A. and Mller N.S. (2011). Discrepancy Analysis of State Sequences. *Sociological Methods & Research 40(3)*, 471–510. `http://smr.sagepub.com/content/40/3/471.full.pdf`.

Wikipedia (2012). tf*idf — Wikipedia, The Free Encyclopedia. [Online; accessed 20-Jun-2012]. `http://en.wikipedia.org/wiki/Tf-idf`.

Some batch self-organizing maps algorithms for interval-valued data

Francisco de A. T. de Carvalho[1,*]

1. Centro de Informatica - CIn/UFPE
*Contact author: fatc@cin.ufpe.bf

Keywords: Self-organizing maps, Interval data, City-Block distances, Hausdorff distances, Symbolic data analysis

The Kohonen Self Organizing Map (SOM) (Kohonen, 1995) is an unsupervised neural network method with a competitive learning strategy which has both clustering and visualization properties. Different from K-means, SOM uses the neighborhood interaction set to approximate lateral neural interaction and discover the topological structure hidden in the data, and in addition to the best matching referent vector (winner), its neighbors on the map are updated, resulting in regions where neurons in the same neighborhood are very similar. It can be considered as an algorithm that maps a high dimensional data space to lattice space which usually has a lower dimension (generally 2) and is called a map. This projection enables a partition of the inputs into "similar" clusters while preserving their topology. The map training can be incremental or batch.

This presentation gives batch SOM algorithms to manage individuals described by interval-valued variables. Interval-valued variables are needed, for example, when an object represents a group of individuals and the variables used to describe it need to assume a value which express the variability inherent to the description of a group. Interval-valued data arise in practical situations such as recording monthly interval temperatures at meteorological stations, daily interval stock prices, etc. Another source of interval-valued data is the aggregation of huge databases into a reduced number of groups, the properties of which are described by interval-valued variables. Therefore, tools for interval-valued data analysis are very much required (Bock and Diday, 2000).

Bock (2002) and D'Urso and De Giovanni (2002) presented incremental SOM algorithms that are able to manage interval-valued data. More recently, batch SOM based on non-adaptive (De Carvalho and Pacifico, 2002; Hajjar and Hamdan, 2011) and adaptive Euclidean (De Carvalho and Pacifico, 2002) distances as well as non-adaptive City-Block (Hajjar and Hamdan, 2011) and Hausdorff (Hajjar, 2011) distances, have been presented.

Badran et al. (2005) give a batch self-organizing approach similar to the K-means algorithm, consisting in a two-step algorithm with an affectation step where all instances are affected to the closest neuron from the grid, and a representation step, where all neurons are updated. This presentation extends Badran et al. (2005) by giving batch SOM algorithms based on adaptive and non-adaptive City-Block and Hausdorff distances, suitable for objects described by interval-valued variables, that, for a fixed epoch, *optimizes* a cost function. Hausdoff distance has been widely used in pattern recognition and computer graphics to measure the dissimilarity of two point sets (Banghe Li and Yuefeng Shen and Bo Li, 2011) For a fixed epoch, these batch SOM algorithms have two (representation and allocation) or three (representation, weighting and allocation) steps aiming to give a partition and a visualization of the data set. These steps are repeated a number of epochs until that a stopping criterion is reached. In this presentation, for each algorithm, it is given the clustering criterion (objective function) and the main steps of the algorithms (the computation of the best prototypes in the representation step, the computation of the best relevance weights of the

variables if there is a weighting step, and the determination of the best partition in the allocation step). The performance, robustness and usefulness of these SOM algorithms are illustrated with real interval-valued data sets.

References

F. Badran, M. Yacoub and S. Thiria (2005). Self-organizing maps and unsupervised classification, in *Neural Networks: methodology and applications, G. Dreyfus, Ed.*, Springer, Berlin et al., pp. 379–442.

H-.H. Bock (2002). Clustering algorithms and Kohonen maps for symbolic data. *Journal of the Japanese Society of Computational Statistics, 15*, 1–13.

H-.H. Bock and E. Diday (2000). *Analysis of Symbolic Data,* Springer, Berlin et al.

F.A.T. De Carvalho and L.D.S. Pacifico (2011). Une version batch de l'algorithme SOM pour des données de type intervalle. In *SFC 2011, XVIIIème Rencontres de la Société Francophone de Classification (Orléans, France)*, pp. 99–102.

P. D'Urso and L. De Giovanni (2011). Midpoint radius self-organizing maps for interval-valued data with telecommunications application. *Applied Soft Computing, 11*, 3877–3886.

T. Kohonen (1995). *Self-Organisation Maps,* Springer, Berlin et al.

C. Hajjar and H. Hamdan (2011). Self-organizing map based on L2 distance for interval-valued data. In *SACI 2011, 6th IEEE International Symposium on Applied Computational Intelligence and Informatics (Timisoara, Romania)*, pp. 317–322.

C. Hajjar and H. Hamdan (2011). Self-organizing map based on City-Block distance for interval-valued data. In *CSDM2 2011, 2nd International Conference on Complex Systems & Design (Paris, France)*, pp. 181–292.

C. Hajjar(2011). Self-organizing map based on hausdorff distance for interval-valued data. In *SMC 2011, The IEEE International Conference on System, Man and Cybernetics (Anchorage, AK, USA)*, pp. 1747–1752.

Banghe Li and Yuefeng Shen and Bo Li (2008). A new algorithm for computing the minimum Hausdorff distance between two point sets on a line under translation. *Information Processing Letters, 106*, 52–58.

Session II

Visualization & Multidimensional Scaling

November 7, 14:30 - 16:30

Two Quantile Visualisation Methods for Symbolic Data

Monique Noirhomme-Fraiture [1,*], **Teh Amouh** [1], **Paula Brito**[2]

1. Faculté d'Informatique - FUNDP, Namur, Belgium
2. Faculdade de Economia & LIAAD-INESC TEC, Universidade do Porto, Porto, Portugal
*Contact author: mno@info.fundp.ac.be

Keywords: Box-Plot, Quantile visualisation, Symbolic Data Analysis

Some authors in Symbolic Data Analysis (SDA) have developed methods based on the knowledge of the quantile distributions associated to the symbolic data. This is, for instance, the case of Ichino (2011) for PCA, Brito & Ichino (2010) for Clustering. To better communicate the results of their analysis, the authors would like to have visualisation tools at their disposal, allowing representing the quantiles of new obtained objects. The aim of this visualisation is to compare different symbolic objects in a same image. In this work, we propose two methods adapted for continuous numerical variables. The Box Plot representation have been proposed by Tukey (1977). This visualisation, commonly used since then, allows representing the three quartiles $Q(0,25), Q(0,50), Q(0,75)$, as well as the extreme values. Tukey introduced this visualisation to make outliers well visible. The minimum and maximum values may also be replaced by $Q(0,05)$ and $Q(0,95)$. In the initial representation, the width of the box does not convey any information. Later on, the authors have used this width to provide information about the sample size or a confidence interval for the median (Mc Gill *et al* (1978)). In 1988 Y. Benjamini presented other extensions of the Box Plot, to represent the probability density (Benjamini (1988)). This density is estimated from the sample. Two variants are proposed, according to whether we represent the density at the quantiles (histplot) or at each point (vaseplot), varying the box width. Finally, Esty and Banfield proposed the Box-Percentile Plot (Esty & Banfield (2003)). This is a vaseplot type figure, where the width at each observed point "provides precise information about the distribution" rather than a density approximation. In fact, between the minimum and the median the method represents the cumulative function at each observed point, and between the median and the maximum represents its complementary, i.e., the difference to 1 of this cumulative function. The percentiles are not clearly identified, but the form of the graphics provides interesting information about the distribution (symmetry or assymmetry, multimodality, uniformity, outliers).

VISUALISATION FOR SDA

We propose two methods for the visualisation of quantiles of symbolic data. The objective is to be able to compare easily several symbolic objects from the point of view of their quantiles (quartiles or deciles). Here, we focus on continuous numerical variables. Both methods are implemented in R.

Visualisation DCS (Symbolic cumulative Diagram) This visualisation is based on a classical approximate cumulative diagram. Let us recall that an approximate cumulative diagram is a continuous piecewise linear curve. The graphic line crosses the points $(L_i, CD(L_i))$ and is linearly interpolated between these points. In the case of DCS, it crosses the points $(Q(\alpha_i), \alpha_i)$. Therefore, the quantiles are read on the horizontal axis, at the breakpoints of the graphics. To simultaneously represent several objects avoiding overlapping, the curves are moved vertically - see Figure 1-(a).

BPS Visualisation (Symbolic Box Plot) The Symbolic Box Plot visualisation takes inspiration from the Box-Percentile Plot. Each object is represented by a sort of icon with an accordion form. We choose an horizontal rather than a vertical representation. The graphics presents a discontinuity at each quantile so that the values indicated on the horizontal axis at each discontinuity are the quantiles (quartiles or deciles) rather than the observed values, as in the Box Percentile Plot.

Coulours and numbers allow identifying the different objects. The median is indicated by a vertical bar - see Figure 1-(b).

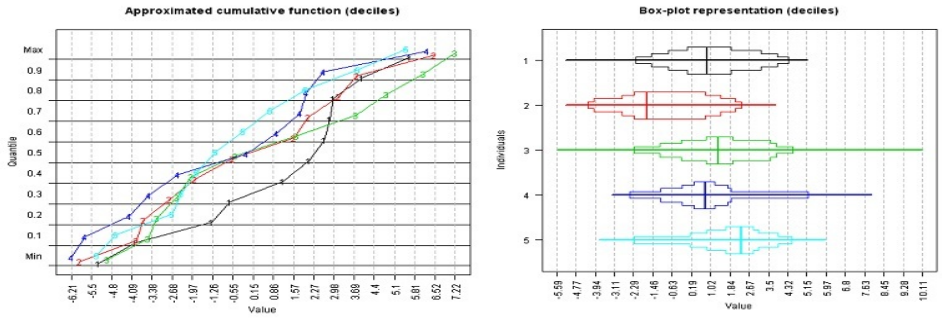


Figure 1: Decile representation of 5 symbolic objects: (a) with DCS; (b) with BPS.

EVALUATION A first evaluation has been set up to establish whether non-experienced users were able to (i) read the quantile values for the symbolic objects and (ii) compare the quantiles of different symbolic objects, using both methods. It was also wished to identify the preferred method and the problems encountered. The two representations were tested with a sample of non-experienced SDA users. The amount of time needed for the task and the number of errors in comparing objects are routhly the same for both methods. On the other hand, the number of reading errors is significantly lower with BPS. As concerns subjective preferences, none of the methods is preferred for comparision and reading tasks, although nine out of ten participants prefer BPS from an esthetical point of view. Extensions of the methods are foreseen as future work.

Acknowledgments

This work is partly funded by the ERDF European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP - 01-0124-FEDER-022701.

References

Benjamini, Y. (1988) Opening the Box of a Boxplot. *The American Statistician, 42 (4)*, 257-262.

Brito, P. and Ichino, M. (2010). Symbolic clustering based on quantile representation. In: *Proc. COMPSTAT 2010*. Paris, France.

Diday, E. and Noirhomme-Fraiture, M. (eds. and co-authors) (2008). *Symbolic Data Analysis and the SODAS Software*. Wiley, Chichester.

Esty, W. and Banfield, J. (2003). The Box-Percentile Plot. *Journal of Statistical Software, American Statistical Association, 8 (1)*.

Ichino, M. (2011). The quantile method for symbolic principal component analysis. *Statistical Analysis and Data Mining*, Wiley. 184-198.

Mc Gill, R., Larsen, W.A. and Tukey, J.W. (1978). Variations of Box Plots. *Journal of the American Statistician, 32*, 12-16.

Noirhomme-Fraiture, M. and Brito, P. (2011). Far beyond the classical data models: Symbolic data analysis. *Statistical Analysis and Data Mining*, *4(2)*, 157–170.

Tukey, J.W., (1977). *Exploratory Data Analysis*.Addison-Wesley, Reading, MA.

Box-plot of symbolic histogram-valued data for data stream analysis

Rosanna Verde[1], **Antonio Irpino**[1], **Lidia Rivoli** [2]

1. Dipartimento di Studi Europei e Mediterranei, Seconda Universitá degli Studi di Napoli
2. Dipartimento di Matematica e Statistica, Universitá degli Studi di Napoli "Federico II"
*Contact author: rosanna.verde@unina2.it

Keywords: Histogram-valued variables, Order statistics, Box-Plot

Symbolic data allow describing typologies and group of observations using set-valued descriptors. A particular type of descriptor is the Histogram Variable. In the literature several proposals have been done for the definition of basic statistics for Histogram Variables (Verde and Irpino, 2007; Arroyo et al., 2011; Rivoli et al., 2012), and also data analysis techniques have been proposed (clustering, regression, dimensionality reduction techniques). In this paper, considering the probabilistic nature of the histogram-valued observations, we propose a method for the construction of a Box-plot for a set of observation describe by a Histogram Variable (i.e., a set of histograms). Starting from the set of the cumulative distribution functions (*cdf*s) associated with a set of histogram data, a Box-plot of *cdf*s is constructed. Using a proposal for the definition of order statistics for histogram-valued data based on the minimization of the $\ell 1$ Wasserstein distance (see Rivoli et al. (2012)), we compute a minimum, a first quartile, a median, a third quartile and a maximum *cdf*. The Box-plot is the visualization of a five-histogram summary (like the Tukey's five number summary (Tukey, 1977)) and we use it for extending the classical definition of box and whiskers plot to data stream analysis.
The proposed Box-plot is used for the evaluation of the evolution of a data stream over time by an innovative exploratory tool. In particular, using non overlapping time windows (Gama and Pinto, 2006) of the same (predefined) time-width, a part of a data stream is summarized by means of a set of histograms. We propose also some measures related to the Box-plot for identifying evolutions in the data stream and for classifying potential outliers.

References

Arroyo, J., Maté, C., Muñoz San Roque, A., González-Rivera, G. (2011). Smoothing Methods for Histogram-valued Time Series. An application to Value-at-Risk. *Statistical Analysis and Data Mining 4 (2)*, 216–228.

Gama, J., Pinto, C. (2006). Discretization from data streams: Applications to histograms and data mining. In *Proceedings of the ACM symposium on Applied computing*, pp. 662–667. ACM, New York USA.

Rivoli, L., Irpino, A., Verde, R. (2012). The median of a set of histogram data. *XLVI meeting of the Italia Statistical Society, June, 2012, Rome, Italy* `http://meetings.sis-statistica.org/index.php/sm/sm2012/paper/view/2194/109`.

Verde, R., Irpino A.,(2007). Dynamic clustering of histogram data: using the right metric. *Selected contributions in data analysis and classification*, Springer, Berlin Heidelberg, , pp. 123–134.

Tukey, John W. (1977). *Exploratory Data Analysis.* Addison-Wesley.

Symbolic Data Analysis of Interval and Histogram Data: an Algorithmic Approach Illustrated on Multidimensional Scaling

Patrick J.F. Groenen[1⋆]

1. Econometric Institute, Erasmus University Rotterdam, The Netherlands
⋆Contact author: groenen@ese.eur.nl

Keywords: Multidimensional scaling, Constrained optimization, Inequality constraints, Iterative Majorization

Interval data and histogram data can be conceived two members of the same class. For multidimensional scaling (MDS), Groenen, Winsberg, Rodriguez, and Diday (2006) proposed the I-Scal algorithm for representing intervals of dissimilarities between pairs of objects by the interval obtained of minimum and maximum distances of the two rectangles representing both objects. This model was extended to handle replications of the interval dissimilarities (for example, interval dissimilarities between all pairs of objects replicated available at T time points) in a three-way version of I-Scal called 3WaySym-Scal by allowing stretching or shrinking of the rectangles per replication, see Groenen and Winsberg (2007). In Groenen and Winsberg (2006), an MDS model for histogram data was proposed by only adding some extra constraints onto the stretching and shrinking of the rectangles.

The core of these three MDS models for symbolic data lies in the minimization of a loss function that models some objective under appropriate constraints to handle interval or histogram data. These constraints generally take the form of inequality constraints. In combination with iterative majorization, the updates satisfying these constraints come down to solving a quadratic program for which efficient procedures exist.

In this presentation, I will set out the main ingredients for modeling interval and symbolic data through least-squares loss function. The approach is illustrated by MDS models for interval and histogram dissimilarities. These models are fitted on empirical data sets and their results discussed.

References

Groenen, P.J.F. & Winsberg, S. (2006). Multidimensional scaling of histogram dissimilarities. In *V. Batagelj, H.-H. Bock, A. Ferligoj, A. Žiberna (Eds.), Data science and classification*, pp. 161–170. Berlin, Springer.

Groenen, P.J.F. & Winsberg, S. (2007). 3WaySym-Scal : Three-Way Symbolic Multidimensional Scaling. In *P. Brito, P. Bertrand, G. Cucumel, F. de Carvalho (Eds.), Selected Contributions in Data Analysis and Classification, Series: Studies in Classification, Data Analysis, and Knowledge Organization*, pp. 55-67.

Groenen, P.J.F., Winsberg, S., Rodriguez, O., & Diday, E. (2006). I-Scal: Multidimensional scaling of interval dissimilarities. *Computational Statistics and Data Analysis 51*, 360–378.

Multidimensional scaling with the non-concentric hypersphere and hyperbox models for percentile-valued dissimilarity data

Yoshikazu Terada[1,⋆], **Hiroshi Yadohisa**[2]

1. Division of Mathematical Science, Graduate School of Engineering Science, Osaka University, Japan
2. Department of Culture and Information Science, Doshisha University, Japan
⋆Contact author: terada@sigmath.es.osaka-u.ac.jp

Keywords: Multidimensional scaling, histogram dissimilarities, fuzzy dissimilarities

Multidimensional scaling (MDS) is one of the important methods for analyzing dissimilarity data. In classical data analysis, each object is represented as a point in $\mathbb{R}^p$. The dissimilarity between two objects is represented as a single value. The classical multidimensional scaling is a method for representing each object as a point in low dimensional space, in such a way as to approximate the given dissimilarities between objects by the distances between points. In symbolic data analysis (SDA), more complex dissimilarity data may appear because higher level objects, referred to concepts, are dealt with. Dissimilarity between two objects in SDA may be described in various ways, including using a single value, an interval, a histogram, and so on. For example, if a large dissimilarity data about individuals (first level objects) is aggregated, dissimilarity data about concepts (second level objects) by using minimum and maximum values of dissimilarities between individuals in each concept, interval-valued dissimilarity data are obtained. It is difficult to analyze such complex dissimilarity data by using classical MDS method without loss of information.

For interval-valued dissimilarity data, Denœux and Masson (2000) proposed the hypersphere and hyperbox models of MDS and use a gradient descent method for solving them. Moreover, MDS for interval-valued dissimilarity data is defined as a method to represent each object as a region in low dimensional space, in such a way as to approximate the given upper and lower dissimilarities between objects by the maximum and minimum distances between regions. Groenen et al. (2006) proposed an improved algorithm based on iterative majorization, called the "I-Scal,h for the hyperbox model. For the hypersphere model, Terada and Yadohisa (2010) proposed the I-Scal algorithm.

In most cases, interval-valued dissimilarity data consists of maximum and minimum values. However, maximum and minimum values are susceptible to the effect of outliers. Thus, it is better to use the percentile-valued dissimilarity data which consists of nested percentile intervals. For histogram-valued (percentile-valued) dissimilarity data, Groenen and Winsberg (2006) proposed the "Hist-Scal" algorithm, which can be considered as an extension of the hyperbox model I-Scal, in that it focuses on quantiles of dissimilarities. For fuzzy dissimilarity data, Masson and Denœux (2002) proposed the similar model, which is considered as an extension of the hypersphere model for interval-valued dissimilarity data. On the Hist-Scal algorithm, the solution does not always improve after each iteration since iterative majorization is used in combination with the weighted monotone regression in each iteration. Terada and Yadohisa (2011) proposed the improved algorithm, called "the concentric hyperbox Percen-Scal algorithm". These models assume that each object is represented by the nested hyperboxes (or hyperspheres) which have a same center point. Such models are called "the concentric hypersphere and hyperbox models." However, the concentric assumption is very strict condition and not natural in most cases.

In this study, a necessary and sufficient condition for that two hyperboxes (two hyperspheres) are nested is derived. New MDS models for percentile-valued dissimilarity data, called "the non-

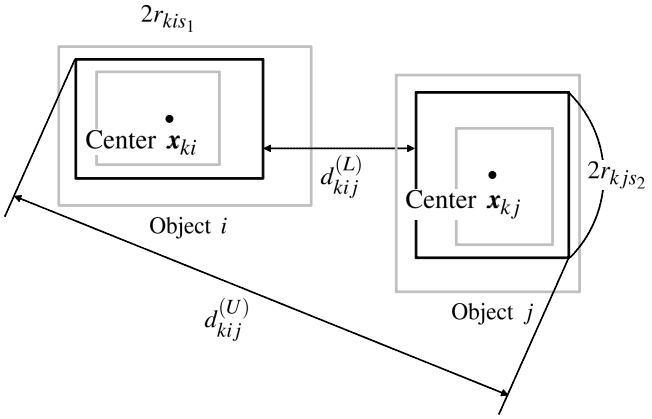


Figure 1: Non-concentric hyoerbox model

concentric hypersphere and hyperbox models," are proposed. In these model, more general nested hyperspheres or hyperboxes, which do not necessarily have a same center point, are used for representing an object (see, Fig. 1). These models can be considered as the natural extension of the models proposed by Masson and Denœux (2002) and Groenen and Winsberg (2006). Moreover, efficient algorithms for these models are proposed based on iterative majorization. The computational costs of proposed algorithms are lower than the BFGS algorithm. Finally, the concentric and non-concentric models are applied to some datasets and the results of these models are compared for establishing the efficiency of new models empirically.

References

Denœux, T., Masson, M. (2000). Multidimensional scaling of interval-valued dissimilarity data, *Pattern Recognition Letters*, **21**, 82–92.

Groenen, P. J. F., Winsberg, S., Rodríguez, O., Diday, E. (2006). I-Scal: Multidimensional scaling of interval dissimilarities, *Computational Statistics & Data Analysis*, **51**, 360–378.

Groenen, P. J. F., Winsberg, S. (2006). Multidimensional scaling of histogram dissimilarities, *In: Batagelj, V., Bock, H. H., Ferligoj, A. and Ziberna, A. (Eds.): Data Science and Classification*, Springer-Verlag, 581–588 (2006).

Masson, M., Denœux, T. (2002) Multidimensional scaling of fuzzy dissimilarity data, *Fuzzy Sets and Systems*, 128, 339–352.

Terada, Y., Yadohisa, H. (2010) Hypersphere model MDS for interval-valued dissimilarity data, *Proceedings of the 27th annual meeting of the Japanese Classification Society*, 35–38 (2010).

Terada, Y., Yadohisa, H. (2011) Multidimensional scaling with hyperbox model for percentile dissimilarities, *In: Watada, J., Phillips-Wren, G., Jain, L. C., and Howlett, R. J. (Eds.): Intelligent Decision Technologies* Springer Verlag, 779–788.

Session III

Applications I

November 7, 17:00 - 18:30

SDA framework is the tool for Big Data Analysis?

Hiroyuki MINAMI[1,*], **Masahiro MIZUTA**[1]

1. Information Initiative Center, Hokkaido University, JAPAN
*Contact author: min@iic.hokudai.ac.jp

Keywords: Massive Data, High-Performance Computing, Cloud system

"Big data" is going to be one of the buzz words in the world. The definition is not clear, but the word is typically used for tons of data which cannot be handled with conventional techniques directly, for example, the size is beyond 10 Tera Bytes. The characteristics are represented by a triple of "*V*", which are ***Volume***, ***Variety*** and ***Velocity***.

Analysts have begun to struggle them and we can read many reports those kinds of the data. However, most are focused on how to handle them in computers, not in statistics. Cloud system and its related technology are powerful tools, but they would just offer us the methodology, without any discussion on the quality from the statistical viewpoint.

As far as a typical user on statistics would conduct "Big data" with the conventional ways, it would take tons of time, or the execution might be corrupted. We have some techniques to reduce the size of a dataset in advance (e.g. sampling, projection in RDB) but they might lose some information and the third "*V*" urges us to handle them dynamically.

We are sure that the framework of Symbolic Data Analysis has potential to find the solution to handle "Big data" properly from both viewpoints of statistics and computer engineering with the following advantages:

Variation To adopt SDA framework, we can handle the targeted big data with Symbolic objects including the variations (e.g. interval value, modal value).

Venture Even if the size were huge, we would have to try many statistical approaches from the viewpoint of exploratory data analysis (EDA). Conventionally it must be hard for huge data, however, we could do it with Symbolic Object.

Validity In SDA framework, we represent the data with *Variation*. It also give us power to keep a statistical property, in contrast to simple data reduction techniques.

Consequently, we propose that SDA for "Big data" represents 3 additional "*V*"s and it is the best solution to handle the data adequately with statistical attention and computer intensive style. We call all "*V*"s "*BIG DATA v6 with SDA*", as an analogy of IPv6 (massive Internet address framework).

It might be tough work to develop SDA analytical system suitable in High Performance Computer Infrastructure. Fortunately, we have an epoch computer cloud system suitable for its development, including Hadoop and MapReduce system in Hokkaido University. In our talk, we discuss the affinity between "Big data" and SDA framework, the system development and practical examples.

References

Rather,B. (2012). *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data (2nd ed.).* CRC Press.

Applications I

Social Networks as Symbolic Data

Giuseppe Giordano [1,⋆], **Paula Brito**[2]

1. Department of Economics and Statistics, Università di Salerno, Italy
2. Fac. Economia & LIAAD INESC TEC, Universidade do Porto, Portugal
⋆Contact author: ggiordan@unisa.it

Keywords: Histogram-Valued Data, Social Network Analysis, Symbolic Data.

Network data refer to a set of actors and their relationships commonly described and represented in the mathematical framework of *Graph Theory*. The graph data structure is characterised by two sets of nodes and edges. Let $\mathcal{G}(N, E)$ be the graph represented by the set *N* of nodes (vertices) with cardinality $n = |N|$ and by the set *E* of edges with cardinality $m = |E|$.

The degree of a node is defined as the number of edges that connect to it. As a starting point we refer to undirected simple finite graphs, that is edges have no orientation, no loops are considered and there is no more than one edge between any two nodes. For such kind of graph a suitable matrix representation is defined through an *adjacency matrix* which is a square symmetric matrix holding all zeros on the main diagonal. The entries in the cells may indicate binary values $0/1$ indicating the absence/presence of the corresponding edges; in such a case, the sum by rows (or by columns) defines the node degree.

The structural analysis of a network is basically performed at descriptive purpose and originated in the framework of *social network analysis* (see Wasserman & Faust, 1994). There exists a large number of natural characterizations of a network based upon degree distribution and many others metrics defined in terms of centrality measures (see Freeman, 1979) for definitions and interpretations). From the most important node-level statistics are *degree*, *closeness*, *betweenness* and *eigenvector centrality*. Global statistics could be computed to capture some topological characteristics of the network and assessing the presence of subgroups (social structures). An example are the indices such as *network density*, the *diameter* of the net, the *number of cliques* and the *size of the largest clique*. We recall that a *clique* in a network represents the set of nodes who have all possible edges present among themselves (Hanneman & Riddle, 2011). The interpretation of such global indices, as well as the statistical distribution of the node characterization is assumed of great interest in many knowledge fields ranging from Marketing to Transport, from Sociology to Economics, from Physics to Medicine, and much more. Topology-related metrics have been found as peculiar features of classes of network. The definition of the graph structure $\mathcal{G}$ as a complex data object should consider the different structural information that can be of interest to retrieve.

The basic idea is to aggregate information attached to each node in terms of its centrality and role in the network and express it as symbolic data by means of interval or histogram-valued variables (see, for instance, Bock & Diday, 2000; Noirhomme-Fraiture & Brito, 2011) so that the whole network could be expressed through the logical union of such different measurements. In this work, we consider the *betweenness centrality distribution, the closeness distribution, the degree distribution, the eigenvector centrality and the edge betweenness centrality distribution*, which are represented, for each network, by histogram-valued variables. Finally, a symbolic data table is built, where each row pertains to a different network and columns to the network indices. That is, each row defines a *Network Symbolic Object (NSO)*. Symbolic data analysis of *NSO* could be applied for the sake of comparisons among several networks emerged at different occasions in time, computing similarities among networks, or representing networks as "points" on a reduced embedding (metric space), to cite just a few possibilities.

In the present study, a simulation study is carried out to generate several artificial network data structures. The traditional network analysis of such data then produces a raw symbolic data table, taking into account the statistical distributions of the main network indices.

Multivariate symbolic data analysis may then be performed on the obtained symbolic data array. In a first step we follow a clustering approach, using different attribute representations, different combinations of attributes and dissimilarity measures. Classical hierarchical clustering, based on a quantile representation (see Ichino, 2008) of the symbolic network data are performed, using different aggregation indices, and provide dendrograms on the set of networks. Other distances more adapted to the type of data at hand - in particular, the Mallow distance (see Verde & Irpino, 2008) - are also used. Conceptual clustering approaches, which take the network symbolic descriptions directly into account provide a different insight. From another point of view, discriminant analysis allows putting in evidence the role of the different retrieved attributes and their discriminant power as relates to the various network classes.

Acknowledgements
This work is partially funded by:
- ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness);
- National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP - 01-0124-FEDER-022701;
- FARB Research Project (2011) "Metodi statistici ad alta complessità di calcolo per lo studio delle reti sociali", Resp. Giuseppe Giordano, University of Salerno. Italy.

References

Bock, H.-H., Diday, E. (2000). *Analysis of Symbolic Data*. Springer, Berlin (2000).

Freeman, L.C. (1979). Centrality in Social Networks I: Conceptual Clarification. *Social Networks, 1*, 215–239.

Hanneman, R. A., Riddle, M. (2011). *Concepts and Measures for Basic Network Analysis*. The Sage Handbook of Social Network Analysis. SAGE.

Ichino, M. (2008). Symbolic PCA for Histogram-Valued Data. In: *Proc. IASC 2008*. Yokohama, Japan.

Noirhomme-Fraiture, M., Brito, P. (2011). Far Beyond the Classical Data Models: Symbolic Data Analysis. *Statistical Analysis and Data Mining, 4 (2)*, 157-170.

Verde, R., Irpino, A. (2008). Comparing Histogram Data Using a Mahalanobis-Wasserstein Distance. In: *Proc. COMPSTAT 2008*, Paula Brito (Ed.), pp. 77-89. Physica-Verlag Heidelberg.

Wasserman, S., Faust, K. (1994). *Social Networks Analysis: Methods and Applications.* Cambridge University Press, New York.

A symbolic solution to a precision agriculture problem

Javier Arroyo[1,*], **María Guijarro**[1], **M. Isabel Riomoros**[2], **Gonzalo Pajares**[1]

1. Departamento de Ingeniera del Software e Inteligencia Artificial. Universidad Complutense de Madrid
2. Departamento de Sistemas Informáticos y Computación. Universidad Complutense de Madrid
*Contact author: javier.arroyo@fdi.ucm.es

Keywords: Computer vision, Histogram data, Image segmentation, Precision agriculture, Texture identification

Precision Agriculture (PA) relies on new technologies to, among other things, match farming practices more closely to crop needs, such as the use of fertilizer or herbicide inputs. The goals include minimizing production costs and avoiding excessive soil pollution during herbicide applications. A common setting is the use of a robot-driven vehicle equipped with computer vision sensors that acts over a site-specific area of a larger farm (Davis et al., 1998). The vehicle automatically applies nutrients or pesticide only to the crop and not to the soil. This setting raises important issues in computer vision, such as the image segmentation problem. Computer vision sensors take images of the field in real time, but how to automatically discriminate between soil and crop?

The main problem is to establish a threshold to each image (Meyer and Neto, 2008). The process to set the threshold described in Guijarro et al. (2011) is shown in Figure 1. The first step is the segmentation of the green color using a green vegetation index such as the excess green index, ExG, (Woebbecke et al., 1995). As a result a greyscale image is obtained. The second step is to apply a thresholding method to set the threshold value of the greyscale range (0-255) that discriminates between crop and soil. Gonzales-Barron and Butler (1995) reviews the more popular thresholding techniques. The result of the technique depends on their assumptions about the content of the image. In the work by Guijarro et al. (2011), the Otsu method (Otsu, 1979) is applied. However, value of the threshold works properly for a specific set of images, but if the conditions change (e.g. illumination or stage of grown of the plants), the threshold changes too. This is a problem in the case of a robot-driven vehicle, because it is not possible to manually set the optimal threshold value depending on the conditions.

This paper proposes the use of symbolic data analysis to overcome this problem. A greyscale image can be represented by its greyscale histogram. The greyscale histogram of an image represents the distribution of the pixels in the image over the grey-level scale. It can be visualised as if each pixel is placed in a bin corresponding to the color intensity of that pixel (see the histogram in Figure 1). This histogram representation of an image is a symbolic data and, consequently, images can be analyzed using symbolic data analysis methods (Billard and Diday, 2006). In this case, it will be used an instance-based learning approach to set the threshold for a given image whose optimal threshold is unknown. The training instances will be the histograms of greyscale images whose optimal threshold value has been manually determined. Given a new image, its threshold will be determined taking into account the threshold values of the closest training examples. These methods are also known as locally weighted learning methods (Atkeson et al., 1997), one of them, the k-nearest neighbors (k-NN) has been successfully applied to deal with histogram data Arroyo and Maté (2009). This paper will compare different strategies based on locally weighted learning methods to automatically binarize canopy crop images, separating soil and crop, that can be used in a robot-driven vehicle in real time.

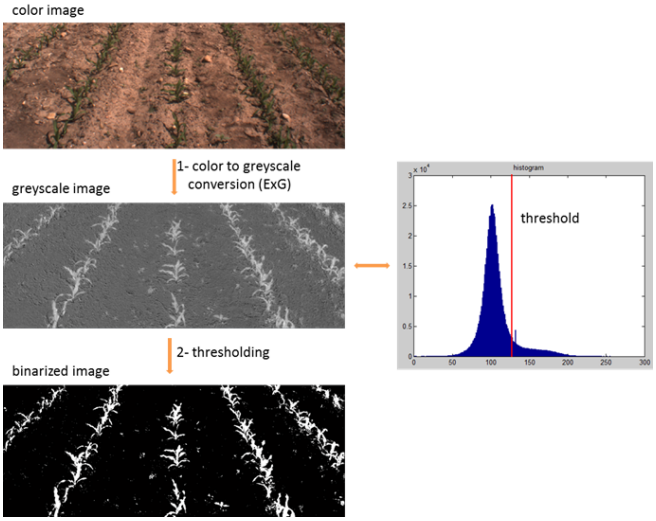


Figure 1: The thresholding process

References

Arroyo, J. and C. Maté (2009). Forecasting histogram time series with k-nearest neighbours methods. *International Journal of Forecasting 25*, 192–207.

Atkeson, C. G., A. W. Moore, and S. Schaal (1997). Locally weighted learning. *Artificial Intelligence Review 11*(1–5), 11–73.

Billard, L. and E. Diday (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Chichester: Wiley & Sons.

Davis, G., W. Casady, and R. Massey (1998). *Precision Agriculture: An Introduction*. Number WQ450 in Water Focus Quality Guide. University Extension, University of Missouri.

Gonzales-Barron, U. and F. Butler (1995). A comparison of seven thresholding techniques with the k-means clustering algorithm for measurement of bread-crumb features by digital image analysis. *Journal of Food Engineering 74*, 268–278.

Guijarro, M., G. Pajares, I. Riomoros, P. J. Herrera, X. P. Burgos-Artizzu, and A. Ribeiro (2011). Automatic segmentation of relevant textures in agricultural images. *Computers and Electronics in Agriculture 75*(1), 75–83.

Meyer, G. E. and J. a. C. Neto (2008). Verification of color vegetation indices for automated crop imaging applications. *Computers and Electronics in Agriculture 63*(2), 282–293.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics 9*(1), 62–66.

Woebbecke, D., G. Meyer, K. Von Bargen, and D. Mortensen (1995). Color indices for weed identification under various soil, residue and lighting conditions. *Transactions of the ASAE 38*, 259–269.

Session IV

Regression

November 8, 9:00 - 11:00

An Overview of Some Regression Models for Interval-valued Symbolic Data

Lynne Billard[1*], Wei Xu[1]

1. University of Georgia
*Contact author: lynne@stat.uga.edu

Keywords: Centers, Ranges, Symbolic moments

Abstract:

We consider some classically based methods for fitting a multiple regression model to interval-valued data (de Carvalho et al., 2004; Lima Neto et al., 2005; Lima Neto and de Carvalho, 2010). Then, a so-called symbolic model is fitted where now the regression parameters are estimated by using the symbolic sample covariance and variance functions of Billard (2008) and Bertrand and Goupil (2000). Also, a min/max function is used to calculate the interval end-points for prediction of the response variable for given predictor variable intervals. To compare methods, a symbolic correlation between the observed and predicted intervals is introduced as a new performance measure. The various methods are compared first using simulated datasets and then an actual dataset.

References

Bertrand, P. and Goupil, F. (2000). Descriptive statistics for symbolic data. In *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data* (eds. H.-H. Bock and E. Diday). Springer-Verlag, Berlin, 103-124.

Billard, L. (2008). Sample covariance functions for complex quantitative data. In *Proceedings, World Conferences International Association of Statistical Computing 2008*, (ed. M. Mituza and J. Nakano). Yokohama, Japan.

de Carvalho F.A.T., Lima Neto, E.A. and Tenorio, C.P. (2004). A new method to fit a linear regression model for interval-valued data. In *Lecture Notes in Computer Science, KI2004 Advances in Artificial Inteligence*. Springer-Verlag, 295-306.

Lima Neto, E.A and de Carvalho F.A.T. (2010). Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics and Data Analysis* 54, 333-347.

Lima Neto, E.A., de Carvalho F.A.T. and Freire, E.S. (2005). Applying constrained linear aggression models to predict interval-valued data. In *Lecture Notes in Computer Science, KI: Advances in Artificial Inteligence* (ed. U. Furbach). Springer-Verlag, Brelin, 92-106.

Lima Neto, E.A., de Carvalho F.A.T. and Tenorio, C.P. (2004). Univariate and multivariate linear regression methods to predict interval-valued features. In *Lecture Notes in Computer Science, AI 2004 Advances in Artificial Intelligence*. Springer-Verlag, Berlin, 526-537.

Regression

New developments in linear regression models with histogram-valued variables

Sónia Dias[1,⋆], **Paula Brito**[2]

1. ESTG, Instituto Politécnico Viana do Castelo & LIAAD–INESC-TEC, Universidade do Porto, Portugal
2. Faculdade de Economia & LIAAD–INESC-TEC, Universidade do Porto, Portugal
⋆Contact author: sdias@estg.ipvc.pt

Keywords: Data with variability, linear regression, histogram-valued variables, quantile functions, Mallows distance.

In the 80's, Schweizer stated that: "distributions are the numbers of the future". Nowadays, many studies may be found that generalize the concepts and methods of classical statistics to "new kinds" of variables, when the observations are sets of values, intervals, distributions. Symbolic Data Analysis plays a crucial role in the development of these studies (Billard and Diday, 2006), (Noirhomme and Brito , 2011).

In recent years, we have been developing a linear regression model applied to data presenting inherent variability, that we named *Distribution and Symmetric Distributions (DSD) Regression Model* (Dias and Brito, 2011). This model allows predicting distributions, represented by their quantile functions, from distributions of explicative variables considering that the relationship may be either direct or inverse. For this kind of variables, this question is most relevant, given that multiplying a quantile function by a negative number does not lead to a non-decreasing function. To solve this problem, our proposal is to include in the linear regression model both the quantile functions $\Psi^{-1}_{X_k(j)}(t)$, that represent the distributions that the explicative histogram-valued variables X_k take for each unit j, and the quantile functions that represent the respective symmetric histograms $-\Psi^{-1}_{X_k(j)}(1-t)$. The predicted quantile function for unit j, is then obtained from

$$\Psi^{-1}_{\widehat{Y}(j)}(t) = \gamma + \alpha_1 \Psi^{-1}_{X_1(j)}(t) - \beta_1 \Psi^{-1}_{X_1(j)}(1-t) + \ldots + \alpha_p \Psi^{-1}_{X_p(j)}(t) - \beta_p \Psi^{-1}_{X_p(j)}(1-t)$$

with $t \in [0,1]\,;\, \alpha_k, \beta_k \geq 0,\, k \in \{1,2,\ldots,p\}$ and $\gamma \in \mathbb{R}$.

As we are working in the semi-vectorial space of the quantile functions, where the defined operations are the addition of the quantile functions and the product of the quantile function by a real positive number, we may consider a quantile function as an independent parameter instead of a real number. This new approach allows the model to be more flexible. In the first situation, when the independent parameter is a real number, it will only influence the fit of the centers of the predicted subintervals of the histogram. This influence will be equal in all subintervals. Considering a quantile function as an independent parameter, it will be estimated to allow predicting quantile functions where the center and half-range of the subintervals of each histogram may be influenced in different ways. For these reasons we may expect better results with this method. Therefore, we propose an extension of the *DSD model* where the predicted quantile function for unit j, is obtained from

$$\Psi^{-1}_{\widehat{Y}(j)}(t) = \Psi^{-1}_{Constant}(t) + \alpha_1 \Psi^{-1}_{X_1(j)}(t) - \beta_1 \Psi^{-1}_{X_1(j)}(1-t) + \ldots + \alpha_p \Psi^{-1}_{X_p(j)}(t) - \beta_p \Psi^{-1}_{X_p(j)}(1-t)$$

with $t \in [0,1]\,;\, \alpha_k, \beta_k \geq 0,\, k \in \{1,2,\ldots,p\}$.

As the interval-valued variables are a particular case of the histogram-valued variables, the *DSD model*, in both versions, may also be applied, to predict intervals from other intervals.

To determine the parameters of the models it is necessary to solve a quadratic optimization problem, subject to non-negativity constraints on the unknowns, and to compute the error measure between the predicted and observed distributions, using the Mallows distance. As in classical analysis, the model is associated with a goodness-of-fit measure whose values range between 0 and 1.

Simulation studies were performed for the proposed models, considering several factors - different distributions for the micro data; different levels of linearity; one or three explicative variables - that allow analyzing the behavior of the estimated parameters and the performance of the model. Results show that the models have similar behavior when all the observations of the histogram-valued explicative variables come from the same distribution (uniform, symmetric, asymmetric) or from a mixture of different distributions. The goodness-of-fit measure reflects well the level of linearity between the variables. In addition, the *DSD models* will also be illustrated with applications to real data tables. The interpretation of the parameters of the *DSD model* will be analyzed based in the real and simulated examples.

In future research, other models and methods in Symbolic Data Analysis based on linear relationships between variables may now be developed using this approach.

Acknowledgments
This work is funded by the ERDF European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP - 01-0124-FEDER-022701.

References

Billard, L., Diday, E., (2006). Symbolic Data Analysis: Conceptual Statistics and Data Mining. John Wiley & Sons, Ltd., Chichester.

Dias, S., Brito, P., (2011). A New Linear Regression Model for Histogram-Valued Variables. In Proceedings of the 58th ISI World Statistics Congress (Dublin, Ireland).

Noirhomme-Fraiture, M. and Brito, P., (2011). Far Beyond the Classical Data Models: Symbolic Data Analysis. *Statistical Analysis and Data Mining 4(2)*, 157-170.

Regression analysis for aggregated symbolic data

Junji Nakano[1*]

1. The Institute of Statistical Mathematics, Japan
*Contact author: nakanoj@ism.ac.jp

Keywords: Multilevel model, Regression analysis, Symbolic data

Symbolic Data Analysis (SDA) handles symbolic data (SD), in which values of a variable can be more complex than the traditional data such as real numbers and categorical values. Typical SD take intervals, histograms or bar charts as variable values, which describe information about the marginal distribution of each variable (Billard and Diday, 2006). SDA provides techniques for handling such SD, including several extensions of regression analysis.

In this paper we use covariance information among variables in each SD for regression analysis of SD. We notice that SD often arise by aggregation of individuals in groups. In this situation, covariance matrices are easily calculated together with traditional SD information, and are naturally used in the regression analysis of SD. We obtain over all regression coefficients which are common to all SD and particular regression coefficients which vary depending on each SD by using penalized least squares estimation. As our motivation is somewhat similar to that of "Multilevel models" (Goldstein, 2011), "Hierarchical linear models" (Raudenbush and Bryk, 2002) or "Linear mixed models" (Jiang, 2007), we discuss the difference between these approaches.

References

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining.* Wiley, West Sussex.

Goldstein, H. (2011). *Multilevel Statistical Models, 4th Edition.* Wiley, West Sussex.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications.* Springer Science+Business Media, New York.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models, Second Edition.* Sage, Thousand Oaks.

Regression

Pattern Classifiers for Interval-valued Data Based on Logistic Regression Models

Alberto P. de Barros[1,2,*], **Francisco de A. T. De Carvalho**[1], **Eufrásio de A. Lima Neto**[3]

1. Centro de Informática da Universidade Federal de Pernambuco
2. Instituto Federal de Educação, Ciência e Tecnologia da Paraíba - IFPB
3. Departamento de Estatística da Universidade Federal da Paraíba

*Contact author: albertopb@ifpb.edu.br

Keywords: Symbolic data analysis, pattern recognition, interval-valued data, logistic regression

The logistic regression models with multi-categorical response (see A. Agresti (2002)) aim to explain a qualitative response variable (nominal or ordinal), consisting of K categories, based on a set of p explanatory variables that can be quantitative, qualitative or both. These models can be applied in many practical situations, such as explaining the preference for a particular car brand, in relation to the other, according to the price of car, age and gender of driver, whether or not a particular accessory, etc. or to classify the performance of an athlete (poor, regular, good, excellent) based on the number of hours of training, type of methodology used by the coach, whether it is nutritional counseling, etc.. These models have been widely used by different companies to predict the level of satisfaction or preference for consumption of their customers.

The purpose of this work is to provide pattern classifiers based on logistic regression models with nominal and ordinal response using interval-valued variables as covariates. Interval-valued variables have been mainly studied in the SDA field, where very often an object represents a group of individuals and the variables used to describe it need to assume a value which express the variability inherent to the description of a group (see H. H. Bock and E. Diday (2000)).

Various methods for the analysis of classical data have been extended to SDA, among which we can cite: factorial analysis, clustering, discriminant analysis, regression analysis, time series analysis, etc. (see M. Noirhomme-Fraiture, and P. Brito (2011). In the area of supervised classification, which is the focus of this paper, A. P. D. Silva and P. Brito (2006) presented three different approaches of supervised learning to interval-valued data based on discriminant analysis. Later, R. M. C. R. Souza, D. C. F. Queiroz and F. J. A. Cysneiros (2011) extended the concept of logistic regression with multi-categoric response for interval-valued data, transforming the response into a set of binary variables.

In this paper, the methodology for classifying a new pattern that was not trained is to assign it to the class with greatest estimated probability by logistic regression model adopted. We considered multinomial logistic regression model when the response variable was nominal and ordinal logistic regression model when the response was ordinal. In order to evaluate the performance of these classifiers, in comparison with the approaches proposed by R. M. C. R. Souza, D. C. F. Queiroz and F. J. A. Cysneiros (2011) and A. P. D. Silva and P. Brito (2006), we consider two synthetic and some real interval-valued data sets. For the synthetic data sets, the measure of accuracy for the classifiers was the average error rate of classification computed in the framework of a Monte Carlo simulation schema, in which a learning and test data sets were randomly selected. For the real data sets, the measure of accuracy for the classifiers was the average error rate of classification computed in the framework of a cross-validation leave-one-out schema.

References

H. H. Bock and E. Diday (2000). *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data* Springer-Verlag, Heidelberg.

A. Agresti (2002). *Categorical Data Analysis* John Wiley & Sons, New Jersey.

L. Billard and E. Diday (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining* Wiley, Chichester.

A. P. D. Silva and P. Brito (2006). Linear Discriminant Analysis for Interval Data. *Computational Statistics 21*, 289–308.

R. M. C. R. Souza, D. C. F. Queiroz and F. J. A. Cysneiros (2011). Logistic regression-based pattern classifiers for symbolic interval data. *Pattern Analysis and Applications 14*, 273–282.

M. Noirhomme-Fraiture, and P. Brito (2011). Far Beyond the Classical Data Models: Symbolic Data Analysis. *Statistical Analysis and Data Mining 4*, 157–170.

Session V

Dimensionality Reduction

November 8, 11:30 - 13:00

Nonlinear canonical analysis for bar chart data tables and interpretation by coherency of metabins and diversity of concepts

Edwin Diday[1]

1. Paris Dauphine University CEREMADE Laboratory.
*Contact author: diday@ceremade.dauphine.fr

Keywords: Principal component, bar chart data table, symbolic data trajectories, metabins.

A barchart data table is characterized by symbolic variables which value for each "concept" (here called "individuals") is a bar chart. Table 1 gives an example of such data table where a sample of people of three regions (Vex, Val, Plai) have answered to three questions: Gender (Man or Woman), Insecurity (priority to Fight Against: Unemployment (FAU) or Juvenile Delinquency (JD) or Drug addict (D)), Death penalty (Yes or No). Each sub variable of such symbolic variable, like the numerical variables M and W for the symbolic variable Gender of bar chart value are called "bins". A metabin is a subset of bins taken among the bins of each bar chart variable. Table 2 is an example of metabin data table where the metabins are called Sicor. For example, S2cor expresses "women who think that fighting against unemployment is the priority and who are against Death penalty". A standard linear canonical analysis of the bar chart variables would produce a graphical representation of the correlation of the bar chart variables inside the correlation circle and a representation of the individual considered as fixed points in the canonical factorial plane. Our approach aims to represent the variation of the individuals by a trajectory inside a PCA and the maximal correlation of the bar chart variables by a good choice of the metabins instead by a good linear combination of the bins like in canonical analysis. This approach is based on the following steps. First, we start with a "trajectories data table" randomly chosen or not, where each row is associated to a couple (individual, metabin) and the columns are associated to each of the bar chart variables. Table 3 (for fixed metabins) and 4 (for fixed individuals) gives two examples of such trajectories data tables issued from the Table 2. Second, we exchange the bins between metabins in order to increase the correlation between the bar chart variables, (thus transformed in numerical variables), in Table 3 or 4. The process can be long as there are $\prod_{j=1}^{p} m_j!$ possible metabins. Therefore, several heuristics can be used (see Diday (2012)) to increase the correlation between the bar chart variables, including by parallel calculus. Ichino (2011) suggests building metabins based on a frequencies order. Another possibility is to build the metabins with the most correlated bins (each one taken in a different bar chart variable) inside the correlation circle of a PCA of Table 1. This has been done and so "cor" is attached to the names of the bar chart variables. The third step is to apply a PCA on table 3 and 4 in order to visualize the two kinds of trajectories. In Figure 1 the PCA of Table 3 with the associated trajectories is represented.

Region	**Gender**		**Insecurity**			**Death Pen.**	
-	**M**	**W**	**fAU**	**JD**	**D**	**Yes**	**No**
Vex	0.8	0.2	0.4	0.5	0.1	0.5	0.5
Val	0.7	0.3	0.5	0.2	0.3	0.4	0.6
Plai	0.3	0.7	0.7	0.1	0.2	0.1	0.9

Table 1. Initial bar chart data table

Region	**S1cor**			**S2cor**			**S3cor**		
	M	**JD**	**Yes**	**W**	**FAU**	**No**	**NU**	**D**	**NU**
Vex	0.8	0.5	0.5	0.2	0.4	0.5	NU	0.1	NU
Val	0.7	0.2	0.4	0.3	0.5	0.6	NU	0.3	NU
Plai	0.3	0.1	0.1	0.7	0.7	0.9	NU	0.2	NU

Table 2. Metabins data table from Table 1.

The fourth step aim is the interpretation of the trajectories of Table 3 and 4. The correlation circle of the PCA of Table 3 or 4 is a useful tool as usual for the axes interpretation. Moreover, from the trajectories of individuals we can induce the "coherency" quality of a metabin. We say that a metabin is "coherent" inside a trajectory of individuals when its bins have a monotonic behavior values on the individuals of this trajectory. In other words, each bin value of a coherent metabin for a given

trajectory increases or decreases monotonically on the individuals of this trajectory. In Figure 1, we can see that all the metabins of Table 2 are coherent. The trajectories of metabins, for fixed Individuals, can be interpreted in term of "diversity" (or "uniformity" at contrary) of individuals which is measured by the similarity between the metabins of the trajectory.

	Gender cor	Insecurity cor	Death pen cor
M JD Yes_path_1_Plai	0.3	0.1	0.1
M JD Yes _path_2_Val	0.7	0.2	0.4
M JD Yes _path_3_Vex	0.8	0.5	0.5
W FAU No_path_1_Vex	0.2	0.4	0.5
W FAU No_path_2_Val	0.3	0.5	0.6
W FAU No_path_3_Plai	0.7	0.7	0.9
NU D NU_path_1_Vex	NU	0.1	NU
NU D NU_path_3_Plai	NU	0.2	NU
NU D NU_path_2_Val	NU	0.3	NU

Table 3. Trajectories of individuals (metabins fixed)

	Gender cor	Insecurity cor	Death pen cor
Vex_path_1_ NU D NU	NU	0.1	NU
Vex_path_2_ W FAU No	0.2	0.4	0.5
Vex_path_3_ M JD Yes	0.8	0.5	0.5
Val_path_1_M JD Yes	0.7	0.2	0.4
Val_path_2_ NU D NU	NU	0.3	NU
Val_path_3_ W FAU No	0.3	0.5	0.6
Plai_path_1_ M JD Yes	0.3	0.1	0.1
Plai_path_2_ NU D NU	NU	0.2	NU
Plai_path_3_ W FAU No	0.7	0.7	0.9

Table 4. Trajectory of metabins (individuals fixed)

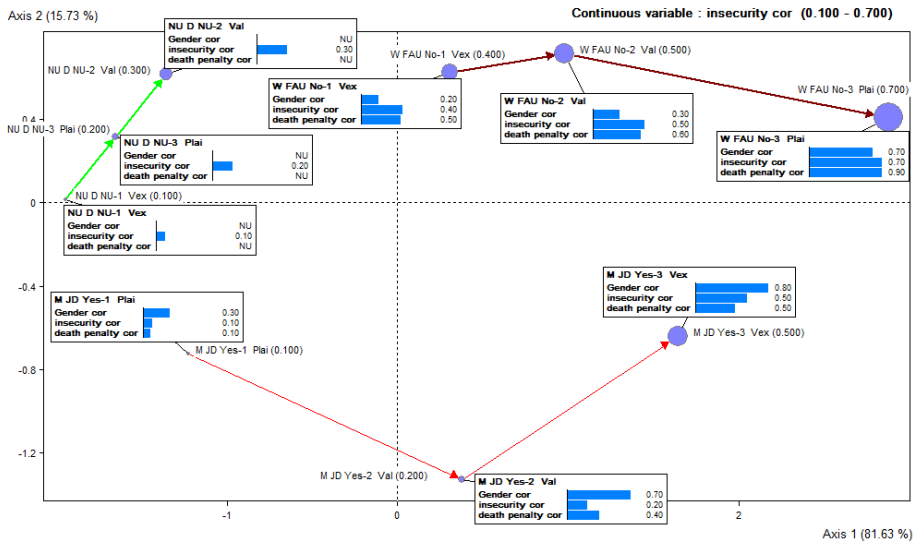


Figure 1. PCA of table Table3 with the trajectories of individuals for fixed metabins. The size of the circles is proportional with the associated bin value.

References

Diday E. (2010) Principal Component Analysis for Categorical Histogram Data: Some Open Directions of Research. *In "Classification and Multivariate Analysis for Complex Data Structures". (B. Fichet, D. Piccolo, R. Verde, M. Vichi eds.).492 pages. Springer Verlag,*

Ichino M. (2011), The quantile method for symbolic principal component analysis, *SAM,Statistical and Data Mining Journal (2011). Volume 4, Issue 2, pages 184–198. Wiley.*

The Duality Problem in Interval Principal Components Analysis

Oldemar Rodríguez[1⋆]

1. CIMPA, School of Mathematics, University of Costa Rica
⋆ oldemar.rodriguez@ucr.ac.cr

Keywords: Symbolic data analysis, interval principal components analysis, correlations circle.

In Cazes (1997) and Billard (2011), the authors proposed the Centers and the Vertices Methods to extend the well known principal components analysis method to a particular kind of symbolic objects characterized by multi–valued variables of interval type. Nevertheless the authors use the classical circle of correlation to represent the variables. The correlation between the variables and the principal components are not symbolic, because they compute the standard correlations between the centers of gravity of variables and the principal components.

It is well known that in standard principal component analysis we may compute the correlation between the variables and the principal components using the duality relations starting from the coordinates of the individuals in the principal plane, also we can compute the coordinates of the individuals in the principal plane using duality relations starting from the correlation between the variables and the principal components. In this paper we propose a new method to compute the symbolic correlation circle using duality relations in the case of interval variables.

References

Billard L. and Diday E. (2006). *Symbolic data analysis: Conceptual statistics and data mining.* Wiley, Chichester.

Billard L., Douzal-Chouakria A. and Diday E. (2011) *Symbolic Principal Components For Interval-Valued Observations*, Statistical Analysis and Data Mining. 4 (2), 229-246. Wiley.

Bock H-H. and Diday E. (eds.) (2000). *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*. Springer, Germany.

Cazes P., Chouakria A., Diday E. et Schektman Y. (1997). *Extension de l'analyse en composantes principales à des données de type intervalle*, Rev. Statistique Appliquée, Vol. XLV Num. 3 pag. 5-24, France.

Chouakria A. (1998) *Extension des méthodes d'analysis factorialle à des données de type intervalle*, Ph.D. Thesis, Paris IX Dauphine University.

Makosso-Kallyth S. and Diday E. (2012) *Adaptation of interval PCA to symbolic histogram variables*, Advances in Data Analysis and Classification July, Volume 6, Issue 2, pp 147-159.

Rodríguez, O. (2000). *Classification et Modèles Linéaires en Analyse des Données Symboliques*. Ph.D. Thesis, Paris IX-Dauphine University.

Discriminant Analysis of Interval Data: Parametric Versus Distance-Based Approaches

A. Pedro Duarte Silva [1,⋆], **Paula Brito**[2]

1. Faculdade de Economia e Gestão & CEGE, Universidade Católica Portuguesa at Porto, Porto, Portugal
2. Faculdade de Economia & LIAAD-INESC Porto LA, Universidade do Porto, Porto, Portugal
⋆Contact author: psilva@porto.ucp.pt

Keywords: Discriminant analysis, Interval data, Parametric modelling of interval data

In this paper, we are interested in the analysis of interval data, i.e., where elements are characterized by variables whose values are intervals on $\mathbb{R}$, and investigate and compare different methods or discriminant analysis of such data.

Distance-based approaches to linear discriminant analysis of interval data are discussed in Duarte Silva & Brito (2006). These approaches lead to representations in the discriminant space in the form of intervals or single points, from which distance-based allocation rules are derived. In Brito & Duarte Silva (2012) a parametric modelling for interval data, assuming multivariate Normal or Skew-Normal distributions for the Midpoints and Log-Ranges of the interval variables, is proposed. The intrinsic nature of the interval variables leads to special structures of the variance-covariance matrix, represented by five different possible configurations. This approach is implemented in an R package, MAINT.DATA Duarte Silva and Brito (2011) (available at the CRAN repository), which includes several tools for modelling and analysing interval data. In particular MAINT.DATA introduces a data class for representing interval data and provides methods and functions for parameter estimation, statistical tests for the different covariance configurations, and parametric discriminant analysis of interval data.

Discriminant analysis of interval data has been investigated by other authors in different contexts. Ishibuchi, Tanaka and Noriko Fukuoka (see Ishibuchi, Tanaka & Fukuoka (1990)) address discriminant analysis of interval data determining interval representations in a discriminant space using a mathematical programming formulation. Approaches of discriminant analysis of interval data based on imprecise probability theory may be found in Nivlet, Fournier & Royer (2001) and Utkin & Coolen (2011). In Lauro, Verde & Palumbo (2000), a generalization of classical Factorial Discriminant Analysis to symbolic data is proposed. This method is based on a numerical analysis of the transformed symbolic data, followed by a symbolic interpretation of the results; it allows considering quantitative, qualitative nominal or modal variables; classification rules are then based on proximities in the factorial plane (see also Lauro, Verde & Irpino (2008)).

This paper evaluates the relative performance of different classification rules for interval data. It compares the distance-based classification rules considered in Duarte Silva & Brito (2006), the parametric classification rules derived from the models discused in Brito & Duarte Silva (2012), and rules proposed by other authors.

Preliminary results show that parametric approaches generally outperform other approaches, and that restricted configurations of the variance-covariance matrix which take into account the particular nature of interval data lead to parsimonious rules, which can be quite effective in reducing expected error rates.

Acknowledgments
This work is partly funded by the ERDF European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT Fundação

para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP - 01-0124-FEDER-022701.

References

Brito, P. and Duarte Silva, A.P. (2012). Modelling interval data with Normal and Skew-Normal distributions. *Journal of Applied Statistics, 39 (1)*, 3-20.

Duarte Silva, A.P. and Brito, P. (2006). Linear discriminant analysis for interval data. *Computational Statistics*, *21 (2)*, 289-308.

Duarte Silva, A.P. and Brito, P. (2011). ***MAINTData:*** *Model and Analze Interval Data*. R package version 0.2, URL http://cran.r-project.org/package=MAINT.Data.

Ishibuchi, H., Tanaka, H. and Fukuoka, N. (1990). Discriminant analysis of multi-dimensional interval data and its application to chemical sensing. *International Journal of General Systems, 16 (4)*, 311-329.

Lauro, N.C., Verde, R. and Palumbo, F. (2000). Factorial discriminant analysis on symbolic objects. In: Bock, H.-H. and Diday, E. (Eds.), *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer, Heidelberg, 212-233.

Lauro, N.C., Verde, R. and Irpino, A. (2008). Factorial discriminant analysis, *in* In: Diday, E. and Noirhomme-Fraiture, M. (Eds.), *Symbolic Data Analysis and the Sodas Software*, Wiley, Chichester, 341-358.

Nivlet, P., Fournier, F. and Royer, J.J. (2001). Interval discriminant analysis: An efficient method to integrate errors in supervised pattern recognition. In: *ISIPTA'01*, 284-292 .

Utkin, L.V. and Coolen, F.P.A. (2011). Interval-valued regression and classification models in the framework of machine learning. In: *7th International Symposium on Imprecise Probability: Theories and Applications*. Innsbruck, Austria.

Session VI

Time Series

November 8, 14:30 - 16:30

The Bayesian Model Averaging approach for interval-valued data

Carlos Maté[1]

1. ETS de Ingeniería (ICAI). Universidad Pontificia Comillas. Alberto Aguilera, 25. Madrid 28015 (SPAIN).
*Contact author: cmate@upcomillas.es

Keywords: Bayesian econometrics, exchange rates, linear regression, interval time series, model uncertainty

Every real world problem or theoretical issue to be considered under the Bayesian statistics framework is generally assuming model uncertainty and the availability of data. For example, these two features are present in all the major areas of a lot of disciplines such as medicine, engineering, economics, and so on. The evolution of Bayesian methods in applied statistics and data analysis in the last 15-20 years is really impressive. According to Bernardo et al. (2008), the application of the Bayesian paradigm saw a spectacular exponential growth during the period 1995-2004, with the number of Bayesian papers in the JCR database rising from 453 to 2254.

At the same time, Bayesian approaches in some statistics communities are not common and this is very surprising. For example, it is very rare to find some papers about Bayesian methods in the symbolic community. In fact, one key issue in the SDA community is: can Bayesian methods be useful with symbolic data? Perhaps the wrong idea is that only when you have solid inferential procedures in an area of statistics you can entry in the Bayesian world. For example, Principal Component Analysis (PCA) is at the core of the beginning of SDA. Recently, Ding et al. (2011) have developed a Bayesian robust PCA. As the title of the paper of Leuenberger and Wegmann (2010) indicates, it is possible to run Bayesian computation and model selection without likelihoods. Hence, if some research team or researcher feels that between different approaches for a real problem with symbolic data some approaches are more adequate than others, they are implicitly assuming a Bayesian focus on their problem.

As Hoeting et al. (1999) state, Bayesian model averaging (BMA) provides a coherent mechanism for accounting for the model uncertainty that is present in all real world situations. Regression analysis is one key method to provide some kind of understanding about such situations where input and output variables can be identified. Model uncertainty arises with regard to the subset of explanatory variables to be included in every model of linear regression to be considered. In addition, according to the experience the research team has, some models will be more likely than others giving a prior distribution about the set of models. That is, they will have some prior idea about ranking the set of models (obviously, this includes the case of no prior idea). Using the information included in the dataset a new ranking for the set of models (a posterior distribution in Bayesian terms) will be obtained whose average will be more effective than the prior ranking in order to explain or forecast the output variables.

Sinova et al. (2012) in a recent paper provide a new and efficient approach to the old problem of doing regression analysis with interval-valued data. However, they do not consider the possibility of model uncertainty in the theoretical development neither the practical cases nor simulations developed.

The foreign exchange market (FOREX) is highly competitive. With respect to its size and importance of the foreign exchange market, Eun and Sabherwal (2002) stated the following. “The market for foreign exchange is the largest financial market in the world. According to the Bank for International Settlements (2001) the average worldwide daily trading in trading foreign exchange markets is

estimated to be US$ 1.2 trillion". More recently, King and Rime (2010) have analyzed the so called "$4 trillion question" giving explanations to the FX growth since 2007. Subsequent to Eun and Sabherwal (2002), and in response to the increasing importance of this market, there has been a number of studies on forecasting exchange rates. The substance of many of these studies is the evaluation of the forecasting performance of one or more exchange rate models. These evaluations suggest that the best predictor of future spot rate is the current spot rate, i.e., the random-walk or "no-change" model, the most commonly used benchmark. This conclusion has also been obtained by Kilian and Taylor (2003). However, using interval-valued data with exchange rates several authors have beaten the random walk (see, for example, Arroyo et al. (2011) and Han et al. (2008)). In addition, others authors like Wright (2008) using BMA have also beaten the random walk with FOREX data in some particular cases.

In this presentation I will show that some nice Bayesian methods can be very useful with symbolic data such as interval-valued data. In particular, considering the findings in Sinova et al. (2012), I will give some guidance about how to implement BMA with regression models for interval-valued data.

Given that both BMA and SDA have separately obtained promising results with FOREX data, the main issue is, would BMA with interval data be able to beat previous approaches? This paper tries to give some answers to this question and concludes proposing some steps for the next future in this new and original approach.

References

Arroyo, J., Espínola, R., Maté, C. (2011). Different approaches to forecast interval time series: A comparison in Finance. *Computational Economics*, 37 (2), 169-191.

Bernardo, J. M.; Arjas, E.; Pilz, J.; Tardella, L.; Robert, C. P.; Wiper, M. P. (2008). European Master in Bayesian Statistics and Decision Analysis. http://www.uv.es/bernardo/BayesEuroMaster.pdf .

Ding, X.ab; He,L.c; Carin, L.c (2011). Bayesian robust principal component analysis. *IEEE Transactions on Image Processing*, 20 (12), 3419-3430.

Eun C. S. and Sabherwal S. (2002). Forecasting exchange rates: Do banks know better? *Global Finance Journal*, **13**, 195-215.

Han, A., Hong, Y., Lai, K.K., Wang, S. (2008). Interval time series analysis with an application to the sterling-dollar exchange rate. *Journal of Systems Science and Complexity*, 21 (4), 550-565.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14 (4), 382-417.

Kilian, L., and Taylor, M. P. (2003). Why is it so difficult to beat the random walk forecast of exchange rates? *Journal of International Economics*, 60 (1), 85-107.

King, M. R. and Rime, D. 2010. The $4 trillion question: what explains FX growth since the 2007 survey? *BIS Quaterly Review*, December, 27-42.

Leuenberger, C., Wegmann, D. (2010). Bayesian computation and model selection without likelihoods. *Genetics*, 184 (1), 243-252.

Sinova, B., Colubi, A., Gil, M.A., González-Rodríguez, G. (2012). Interval arithmetic-based simple linear regression between interval data: Discussion and sensitivity analysis on the choice of the metric. *Information Sciences*, 199, 109-124.

Wright, J. H. (2008). Bayesian Model Averaging and exchange rate forecasts. *Journal of Econometrics*, 146, 329-341.

Multiple Beanplots GCCA in a Temporal Framework

Carlo Drago, Carlo Natale Lauro and Germana Scepi[1,2,⋆]

1. Department of Mathematics and Statistics University of Napoli "Federico II"
⋆Contact author: carlo.drago@unina.it

Keywords: Beanplots, Time Series, Finite Mixture Modeling, Generalized, Canonical Correspondence Analysis

Advances in computer technology have made large data sets of financial data increasingly common and have caused the data frequency to be expanded by containing tick by tick observations. In this context, high frequency data contains the characteristics of all the market transactions. In particular, high frequency data are collected daily on a finer time scale and are typically irregularly spaced. In these cases, problems arise when the data need to be synthesized using some aggregation function. In fact, in this case there is a clear loss of information on the variability of the data in the time interval considered. As well as the SDA approach, various proposals exist in the representation of such data types: interval valued data; boxplot and histograms time series; and more recently beanplot time series- in which the observations are represented as density functions. In this sense, Drago, Lauro, Scepi (2011) proposed the use of the beanplot time series assuming the decomposition of original data as the sum of a model plus an error, to obtain density models, based on a mixture of distributions. The beanplot parameters allow us to synthesize and describe correctly the original data, solving at the same time the problem of massive data storage. A beanplot time series represents a powerful tool to visualize the dynamics and change of patterns of massive data. The corresponding vector time series of the beanplot parameters obtained in the parameterization process is suitable for forecasting and identifying structural change by means of constrained clustering technique (Drago, Lauro, Scepi 2011). In the present contribution we extend our previous approach for single beanplot analysis to the case of multiple beanplot time series. For example, by considering the multiple beanplot time series related to a market the resultant synthesis will be a beanplot representing the entire market (as an index of the entire market, for example, FTSE MIB for the Italian Case). In particular, we propose a new approach to the data analysis in a temporal framework based on using the generalized canonical correlation analysis (GCCA) of the beanplot models over time. The proposed procedure, aims to obtain a synthesis of the multiple beanplot time series represented by a correspondent multi-vector of parameters in a single beanplot time series representing the whole market evolution. Each parameter of the beanplot multiple time series is grouped in a suitable partitioned matrix of homogeneous parameters GCCA allowing us to build a synthesis of the multiple beanplot time series represented by its parameters over time. By using these methods the multiple beanplot time series can be synthesized in a beanplot time series which represents the relevant common information of the original time series. Each canonical component identify the canonical components, associated with the initial multiple parameters. The canonical components can be used for forecasting aims or to change point analysis as we did for the single beanplot time series. The first factor associated to each group of parameters is typically a size factor whereas the second one, if it expresses a significant component, typically reveals cyclical aspects. A representation of both components in the two sides of a single beanplot has high interpretative capacity.

References

Billard, L., Diday, E., (2003). From the statistics of data to the statistics of knowledge: Symbolic data analysis. *J. Amer. Statist. Assoc.* 98 (462), 470–487.

Drago C. Lauro C.N. Scepi G. (2011). Beanplot Data Analysis in a Temporal Framework, In *CLADAG 2011* Pavia.

Engle, R.F., and Sun, Z. (2005) Forecasting volatility using tick by tick data. Technical Report, New York University

Fischer, B., Roth, V., Buhmann, J. M. (2007) Time-series alignment by non-negative multiple generalized canonical correlation analysis. *BMC Bioinformatics*, 8 Suppl 10, S4.

Velden, M. van de and Takane, Y.(2009) Generalized canonical correlation analysis with missing values No EI 2009-28, *Econometric Institute Report*, Erasmus University Rotterdam, Econometric Institute

Yan B., and Zivot E. (2003) Analysis of High-Frequency Financial Data with S-Plus. Technical Report

Modeling and Forecasting Interval Time Series with Threshold Models: An Application to S&P500 Index Returns

Paulo Rodrigues [1], Nazarii Salish[2,*]

1. Banco de Portugal, NOVA School of Business and Economics, CEFAGE
2. University of Bonn, BGSE
*Contact author: salish@uni-bonn.de

Keywords: Interval Time Series, Forecasting, Threshold Models, Forecast Accuracy Measures

Modeling and forecasting interval-valued time series (ITS, hereafter) has received considerable attention in the recent literature. ITS has been introduced as a new field related to multivariate analysis and pattern recognition, where the most influential methodologies for the analysis of integer valued data are exponential smoothing methods, pattern recognition and multivariate models (See e.g., Muñoz et al., 2007, Maia and Carvalho, 2011, Arroyo et al., 2011).

In economic and financial settings, one may consider data registered in an almost continuous way, such as, exchange rate fluctuations, stock prices and returns, and electricity demand. For instance, the daily or weekly highest and lowest prices of assets may be regarded as boundary values of an interval and therefore ITS modelling and forecasting techniques, such as the ones presented in this paper may provide useful tools for the analysis (See also Han et al., 2008, He and Hu, 2009, García-Ascanio and Maté, 2010).

The limited economic and financial application of linear (interval) models when interest lies in the analysis of regime dependence or asymmetric behaviour of the series over the business cycle has lead to the development of a large number of nonlinear models. One class of nonlinear models that has proven to be successful in the literature are the threshold autoregressive (TAR, hereafter) models. For instance, Tong (1990) developed TAR models and applied them to predict stock price movements. Henry et al. (2001) present evidence of threshold nonlinearity in the Australian real exchange rate, and Duaker et al. (2007) develop a contemporaneous TAR model for the bonds market. In the context of ITS several papers have also stressed the importance of nonlinearities (see, e.g., Muñoz et al., 2007, Maia et al., 2008, Maia and Carvalho, 2011). These studies present evidence of the low accuracy of linear approaches to forecast ITS with nonlinear characteristics and introduce the application of neural networks and hybrid methods of forecasting. However, these procedures aim at producing forecasts without explicitly modelling the nonlinear characteristics of the data and to the best of our knowledge there have been no studies in the ITS context that attempt to model and explain nonlinear features. Furthermore, no empirical results on regime dependent ITS forecasts are available so far.

In this paper, econometric methods for regime switching threshold models are adapted to ITS characterized by their center and radius. Since accuracy measures play an important role in the context of ITS analysis we also discuss interval quality measures in the line with related literature (see e.g., Ichino and Yaguchi, 1994, Arroyo and Maté, 2006), as well as additional

forecast descriptive statistics (such as efficiency and coverage rates) in order to provide more information for an objective decision regarding the interval forecast performance. To illustrate the proposed approach, we report an application to a weekly sample of S&P500 index returns. The results obtained are encouraging and compare very favourably to available procedures.

References

Arroyo, J., Espínola, R., and Maté, C. (2011). Different Approaches to Forecast Interval Time Series: A Comparison in Finance. *Computational Economics, 37 (2)*, 169-191.

Arroyo, J. and C. Maté (2006). Introducing interval time series: accuracy measures. *In COMPSTAT 2006, Proceeding in Computational statistics, Heidelberg*, 1139-1146.

Dueker, M., S. Martin and F. Spangnolo (2007). Contemporaneous Threshold Autoregressive Models: Estimation, Testing and Forecasting. *Journal of Econometrics 141*, 517-547.

García-Ascanio, C.; Maté, C. (2010). Electric power demand forecasting using interval time series: A comparison between VAR and iMLP. *Energy Policy 38*, 715-725.

Han, A., Hong, Y., Lai, K.K., Wang, S. (2008). Interval time series analysis with an application to the sterling-dollar exchange rate. *Journal of Systems Science and Complexity, 21 (4)*, 550-565.

He, L.T. and C. Hu (2009). Impacts of Interval Computing on Stock Market Variability Forecasting. *Computational Economics 33*, 263-276.

Henry, Ó., N. Olekaln and P.M. Summers (2001). Exchange Rate Instability: a Threshold Autoregressive Approach. *Economic Record 77*, 160-166.

Ichino, M. and H. Yaguchi (1994). Generalized Minkowski metrics for mixed and feature-type data analysis. *IEEE Trans. on Systems, Man and Cybernetics, 24(1)*, 698-708.

Maia, A.L.S. and F.d.A.T. de Carvalho (2011). Holt's exponential smoothing and neural network models for forecasting interval-valued time series. *International Journal of Forecasting 27*, 740-759.

Maia, A.L.S. and F.d.A.T. de Carvalho, and T.B. Ludermira (2008). Forecasting models for interval-valued time series. *Neurocomputing 71*, 3344-3352.

Muñoz San Roque, A., C. Maté, J. Arroyo and Sarabia, Á. (2007). iMLP: Applying Multi-Layer Perceptrons to Interval-Valued Data. *Neural Processing Letters 25*, 157-169.

Tong, H. (1990). Non-Linear Time Series: A Dynamic System Approach. *New York: Oxford University Press.*

Forecasting bar diagram-valued time series with exponential smoothing methods

Carlos A.G. de Araújo Júnior[1,2,*], **Fracisco de A.T. de Carvalho**[1], **André L.S. Maia**[3]

1. Centro de Informática, Universidade Federal de Pernambuco-UFPE
2. Instituto de Pesquisas Maurício de Nassau-IPMN
3. Centro Regional da Bahia, Fundacentro
*Contact author: cagaj@cin.ufpe.br

Keywords: Time series forecast, exponential smoothing, bar diagram-valued data

In time series analysis the data usually considered are a sequence of data points. Time series where observations are single values are suitable for representing many practical situations. However, they do not faithfully describe phenomena where a set of realizations of the observed variable is available for each time point. In this case variables assume sets of categories or intervals, possibly even with frequencies or weights. This kind of data have been considered in the field of *symbolic data analysis* (SDA), see Bock and Diday (2000) and Billard and Diday (2006). When a set of categories with related frequencies of the observed variable is available for each time point we have a bar diagram-valued time series. This paper introduces two exponential smoothing methods based on simple and Holt's exponential smoothing method to forecast bar diagram-valued time series data. The proposed methods are inspired in the approach introduced by Maia and De Carvalho (2011) to deal with interval-valued time series. The smoothing parameters are estimated by using techniques for non-linear optimization problems with bound constraints. The results are discussed based on two well-known classical performance measurements, which have been adapted here for this particular type of data: the Theil's U statistics and average relative variance (ARV) in the framework of a Monte Carlo experiment. The synthetic data sets take into account different aspects, e.g., sample size and forecast horizons among others. Applications using real bar diagram-valued time series also were considered to demonstrate the practicality of the methods. We considered four examples of real data time series from different contexts. The results demonstrate that the proposed approaches are useful in forecasting bar diagram-valued times series.

References

Arroyo, J. and Maté, C. (2009). Forecasting histogram time series with k-nearest neighbours methods. *International Journal of Forecasting 25*, 192–207.

Billard, L. and Diday, E. (2006). *Symbolic data analysis. Conceptual statistics and data mining* Wiley, Chichester.

Bock, H.H. and Diday, E. (2000). *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data* Springer-Verlag, Heidelberg.

Holt, C.C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting 20*, 5–13.

Maia, A.L.S. and De Carvalho, F.A.T. (2011). Holt's exponential smoothing and neural network models for forecasting interval-valued time series. *International Journal of Forecasting 27*, 740–759.

Session VII

Software

November 8, 17:00 - 18:00

A Symbolic Database for TIMSS

Chiun-How Kao[1,2] , Chih-Wen Ou-Yang[2], Yin-Jing Tien[2], Chuan-kai Yang[1], Chun-houh Chen[2,*]

1. Department of Information Management, National Taiwan University of Science and Technology, Taiwan
2. Institute of Statistical Science, Academia Sinica, Taiwan
*Contact author: cchen@stat.sinica.edu.tw

Keywords: Database, Generalized association plots (GAP), Matrix visualization, SDA, TIMSS

Well organized datasets collected from compelling biomedical experiments or social surveys are critical for demonstrating strength and novelty of newly proposed statistical theories and methodologies. It is essential for the symbolic data analysis (SDA) community to have good databases for exercising new methods and software developed for various types of SDA data. In this study we plan to construct a symbolic database for the TIMSS (Trends in International Mathematics and Science Study) project.

Conducted every four years at the fourth and eighth grades, the TIMSS project provides: (1) a comparative study for improving teaching and learning in mathematics and science for students around the world; (2) related data sets about trends in mathematics and science achievement over time. In this study we shall use datasets from TIMSS (2007) for constructing an SDA database with various level of concepts (country, school, class) and different types of SDA data (interval, (modal) multi-valued, etc.) for each of the two grades (fourth, eighth) and two subjects (mathematics, science) with thirteen test booklets each. Many covariates describing related concepts at different levels will also be included in this database. (Fig. 1)

The TIMSS_SDA database alone with a *R* interface developed throughout this project will be opened to public use. We shall use the GAP (Generalized Association Plots) matrix visualization environment (Chen, 2002; Wu et al. 2010) to introduce related data structure of TIMSS_SDA in our presentation.

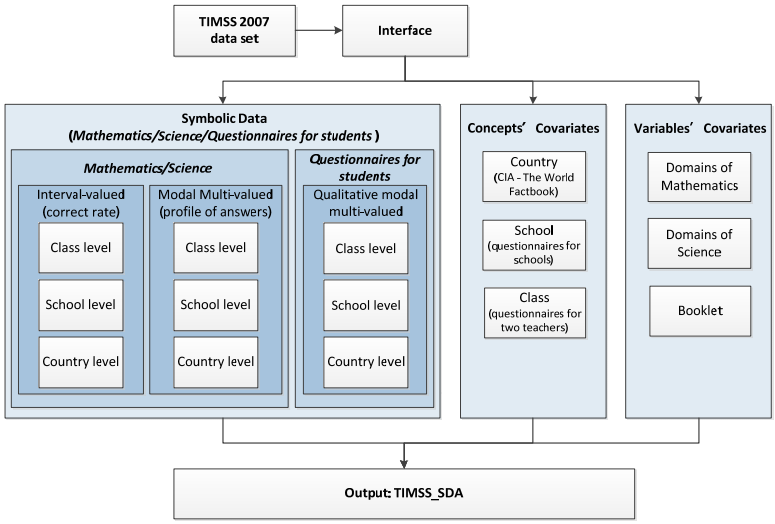


Figure 1. Flowchart of TIMSS_SDA.

References

Chen, C.H. (2002). Generalized association plots: Information visualization via iteratively generated correlation matrices. *Statistica Sinica* 12, 7-29.

Wu H.M., Tien Y.J., and Chen C.H. (2010). GAP: a graphical environment for matrix visualization and cluster analysis. *Computational Statistics and Data Analysis*, 54, 767-778.

TIMSS (2007). TIMSS 2007 Assessment Framework. Copyright © 2009 *International Association for the Evaluation of Educational Achievement (IEA)*. Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Latest developments of the SYR software for Symbolic Data Analysis of complex data

Filipe Afonso[1], Raja Haddad[1,2], Edwin Diday[2]

1. SYROKKO Company, Aéropôle - Bat. Aéronef, 95731 Roissy CDG Cedex, France
2. Paris Dauphine University, 75775 Paris Cedex 16, France
*Contact author: afonso@syrokko.com

Keywords: Symbolic data analysis software, Decision trees, Clustering, Principal component analysis, Visualizing symbolic data.

We present the main evolutions of the SYR software for Symbolic Data Analysis of Complex Data (Afonso et al., 2012).

The SYR software is a SYROKKO company product. Its aim is to extract, from a data file, up to several millions of units, a reduced number of units called "concepts" which summarize the initial data. These units are described by standard categorical or numerical variables, as well as by interval variables, multi-valued variables and by bar-chart and histogram-valued variables. These new kinds of variables allow keeping the internal variation of each concept. The presentation focus on the recent developments extended to any kind of symbolic data:

- Decision trees;

- Principal Component Analysis (PCA);

- New statistics and user-friendly graphical tools for PCA factorial plane and correlation circles, specific to symbolic data;

- Visualization of partition or overlap clustering in PCA factorial planes;

- Visualization of time series, trajectories and pathways in PCA factorial planes.

The software is presented through industrial applications (for example, Courtois et al., 2012). Comparisons with the academic SODAS software for symbolic data (Diday and Noirhomme, 2008) are also given.

Finally, further researches and development of the software are discussed: creation of informative symbolic data, metabins and trajectories of metabins, learning techniques with decision trees.

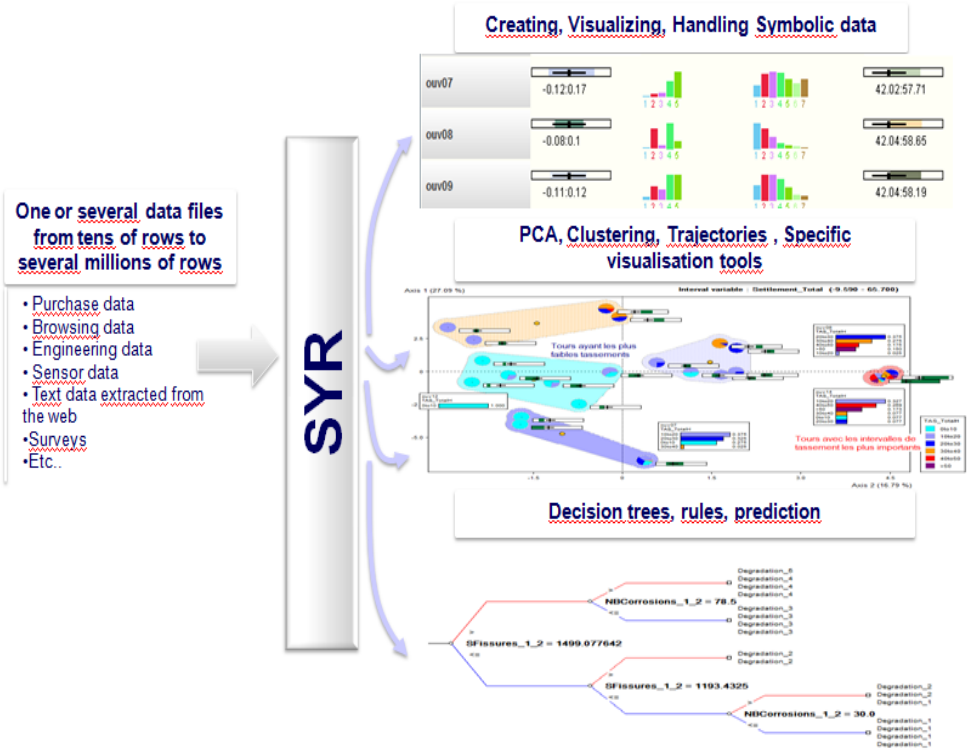


Figure 1 : Overall structure of the SYR software

References

Afonso, F., Haddad, R., Toque, C., Eliezer E.-S., Diday, E. (2012). *User Manual of the SYR Software*, Syrokko internal publication, 70p., http:// www.syrokko.com

Courtois, A., Genest, G., Afonso, F., Diday, E., Orcesi, A., (2012) *In service inspection of reinforced concrete cooling towers – EDF's feedback* ,IALCCE 2012, Viena, Austria.

Diday, E. and Noirhomme-Fraiture, M. (eds and co-authors) (2008). *Symbolic Data Analysis and the SODAS software*. Wiley, Chichester, ISBN 978-0-470-01883-5.

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. 321 pages. Wiley series in computational statistics. Wiley, Chichester, ISBN 0-470-09016-2.

Session VIII

Applications II

November 9, 9:00 - 11:00

Study on Radiation Therapy with Distribution Valued Data

Masahiro Mizuta

Advanced Data Science Laboratory, Information Initiative Center, Hokkaido University, Japan
*Contact author: mizuta@iic.hokudai.ac.jp

Keywords: Dose-Volume Histogram, Selection of Fraction Regimen, Dose Distribution

There are the most common types of cancer treatment, such as surgery, chemotherapy, radiation therapy, and many others. Among them, radiation therapy is in itself painless and with low burden to patients. It is not unusual radiotherapy leads to better treatment outcomes than other treatments. Dose-volume histogram (*DVH*) plays a key role in radiotherapy (radiation therapy) and is a clinically relevant criterion to evaluate a treatment plan quality. The differential DVH is the density function. That means we can regard DVH as a distribution valued data or symbolic description.

The principle of Radiotherapy is to kill cancer cells and minimize damage effect on organs at risk (OAR). Various altered fractionation regimens have been proposed to improve tumor control without increasing late toxicity to normal tissue. Recently, the author has proposed a simple mathematical method to compare conventional multi-fractionated irradiation and hypo-fractionated irradiation based on the LQ model in terms of minimizing radiation damage to OAR under the condition that the effect to the tumor tissue was fixed (Mizuta(2012)). In this paper, we extend the method for 3D dose distribution from the viewpoint of SDA.

Distribution Valued Data
Objects in conventional data analysis are assumed that they are described with a set of numbers with specific structure, *e.g.* a set of multidimensional vectors. But, we, statisticians must analyze a complex data or big data now. It is difficult to describe these kinds of data as a set of vectors. In order to overcome this problem, SDA supplies various data descriptions; interval valued data, modal interval data, categorical data, distribution valued data, *etc*. Distribution valued data is fruitful because of its ability of expression. Professor Diday quoted Schweizer as saying "distributions are the number of the future".

LQ model
There are many models in the radiation survival responses of human tumor cells including Linear-Quadratic model (*LQ model*). The LQ model is commonly used to evaluate and compare different fractionation schedules in radiotherapy. The basic assumption in this study relies on the LQ model for both tumors and normal tissues; the formula $E(d) = \alpha d + \beta d^2$ is used for the effect as a function of absorbed dose d, where α and β are parameters. We can regard $exp(-E(d))$ as survival rate. We use the notations α_1 and β_1 for the tumor and α_0 and β_0 for the OAR as the parameters, respectively. In general, these parameters satisfy $\frac{\alpha_0}{\beta_0} < \frac{\alpha_1}{\beta_1}$.

Radiation Effect on Tumor
For multifractionated radiation therapy with n-fraction dose d, the radiation effect on the tumor is represented by

$$\sum_{i=1}^{n} n(\alpha_1 d + \beta_1 d^2) \tag{1}$$

and fixed as E_1, *i.e.* the survival rate of tumor is $\exp(-E_1)$ *e.g.* 10^{-5}.

Damage Effect on OAR

It should be reasonable to consider that the dose for the OAR is proportional to the dose for the tumor, that is, the dose for the OAR is given by $\delta \times d$, where the dose for the tumor is d and the proportionality factor δ satisfies $0 < \delta$.

If we assume that the δ is constant on OAR, the damage effect on OAR is represented by

$$E_0(d, n) = n(\alpha_0 \delta d + \beta_0 (\delta d)^2). \tag{2}$$

If we assume that the δ is not constant on OAR and the density function *i.e.* differential DVH, is $f_i(\delta)$, the damage effect on OAR is

$$E_0(d, n)(f_i) = -ln \int_0^\infty exp(-n(\alpha_0 \delta d + \beta_0 (\delta d)^2)) f_i(\delta) d\delta, \tag{3}$$

where i is an index of plans of therapy ($i = 1, \cdots, k$). This formulation shows that the damage effect on OAR is determined by n, d and f_i. It is not difficult to solve the constraint optimization problem for each i. We can interpret $E_0(d, n)$ as *functional* or *operator*; the feasible range of the effect on Tumor versus the damage effect on OAR corresponds a function. Functional and operator are important mathematical tools for analysis of distribution valued data.

Concluding Remarks

We discussed radiotherapy for an important example of distribution valued data. An invited session "Analysis of Distributional Data" is approved by SPC of ISI2013 in Hong Kong. I will attend the session with great expectations.

In this paper, we assume that the conventional LQ model (1) for tumor. But a characteristic feature of tumor is repopulation. The LQ model extended for tumor repopulation can be used for this discussion. The results based on the extended LQ model was reported in Mizuta(2012).

References

H. Shirato, M. Mizuta, K. Miyasaka(1995) A mathematical model of the volume effect which postulates cell migration from unirradiated tissues, *Radiotherapy and Oncology* 35, 227–231.

M. Mizuta, S. Takao, H. Date, N. Kishimoto, H. Shirato(2011) Theoretical Comparison between Availabilities of Single- and Fractionated- Irradiation Therapies ASTRO's (American Society for Radiation Oncology) 53rd Annual Meeting, October 2 - 6, 2011. *International Journal of Radiation Oncology Biology Physics*, Vol.81, Number 2, Supplement, P.728.

M. Mizuta, S. Takao, H. Date, N. Kishimoto, K. L. Sutherland, R. Onimaru, H. Shirato(2012). A Mathematical Study to Select Fractionation Regimen based on Physical Dose Distribution and the Linear-Quadratic Model, *International Journal of Radiation Oncology, Biology, Physics*, In press.

M. Mizuta(2012). Optimization of Dose Fractionation based on 3D dose distribution and LQ model, *31st Sapporo International Cancer Symposium 2012 Advanced Radiation Therapy and Cancer Research Up-to-Date*.

A Collaborative Filtering Algorithm based on Histograms Principal Component Analysis

Juan de Dios Murillo[1*] **and Oldemar Rodríguez**[2**]

1. School of Informatics, National University, Costa Rica
2. CIMPA, School of Mathematics, University of Costa Rica
Contact author: * jmurillo@una.ac.cr **, oldemar.rodriguez@ucr.ac.cr

Keywords: Collaborative filtering, recommender systems, symbolic data analysis, histograms principal components analysis.

The collaborative filtering recommender systems (CF) have become an important tool to cope with the information overload problem by acquiring information about the user behavior. In order to increase the user satisfaction companies try to predict their preference based on the user behavior. Recommendation systems are implemented in commercial and non-profit web sites to predict the user preferences, the main functions of them include analyzing user data and extracting useful information for further predictions. Recommendation systems apply very different data analysis techniques to determine the similarity among thousands or even millions of data.

In this paper, we propose a new approach for the collaborative filtering recommender system using Histograms Principal Component Analysis method (HPCA Rodriguez (2000) and Diday (2000)). The main idea of the new algorithm is to use symbolic objects representation with histograms-valued variables to get a dimensionality reduction by applying HPCA, then make the recommendation for a user using clustering methods in symbolic data analysis (Carvalho (2006)). Finally we compare the results with a traditional Eigentaste method proposed by Goldberg (2000). Eigentaste is a collaborative filtering algorithm that uses universal queries to elicit real-valued user ratings on a common set of items and applies principal component analysis (PCA) to the resulting dense subset of the ratings matrix. PCA facilitates dimensionality reduction for offline clustering of users and rapid computation of recommendations. We make this comparison based in data sets for example: MovieLends, Jester and Each Movie.

References

Billard, L and Diday, E. (2003). *From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis*. Journal of the American of the Statistical Association, USA.

Bezerra B. and Carvalho F. (2004). *A symbolic approach for content-based information filtering*. Information Processing Letters, Volume 92, Issue 1, 16 October 2004, Pages 45-52

Carvalho F., Souza R.,Chavent M., and Lechevallier Y. (2006) *Adaptive Hausdorff distances and dynamic clustering of symbolic interval data.* Pattern Recognition Letters Volume 27, Issue 3, February 2006, Pages 167-179

Diday, E., Rodríguez O. and Winberg S. (2000). *Generalization of the Principal Components Analysis to Histogram Data*, 4th European Conference on Principles and Practice of Knowledge Discovery in Data Bases, September 12-16, 2000, Lyon, France.

Goldberg K., Roeder T., et all. (2000). *Eigentaste: A Constant time Collaborative Filtering Algorithm*. IEOR and EECS Departmets, University of California, Berkeley, USA. `http://goldberg.berkeley.edu/jester-data/`.

Herlocker, J.L., Konstan, J.A., Terveen, L.G., and Riedl, J. (2004). *Evaluating Collaborative Filtering Recommender Systems*. ACM. Transactions on Information Systems, USA.

Manolis, G and Konstantinos G. (2006). *A Recommender System using Principal Component Analysis*. Pararell Distributed Procesing Laboratory, Department Applied Informatics. University of Macedonia, Greece.

Rodríguez, O. (2000). *Classification et Modèles Linéaires en Analyse des Données Symboliques*. Ph.D. Thesis, Université Paris IX-Dauphine.

Analyzing European social survey data using symbolic data methods and Syrokko software

Filipe Afonso[1*] , Seppo Laaksonen[2]

1. SYROKKO
2. University of Helsinki
*Contact author: afonso@syrokko.com

Keywords: SDA software, Social survey data, Visualization of symbolic data

We have data from social surveys carried out among European inhabitants. In this study, we are not interested in studying the people themselves but in the comparison of different European countries, or different regions of Europe (Western Europe, Eastern, Nothern, …), or some groups of inhabitants by age, gender, etc... Thus, to study the different European countries or regions or countries x age…, we may describe each of them by all the results of its inhabitants by keeping the within variation of these results. Symbolic Data Analysis (SDA) is proving to aggregate up micro data (at the level of the inhabitants) to higher level units (at the level of the countries or European regions), using symbolic histogram or interval-valued variables. The aggregation is performed by the *SYR* software for SDA from *Syrokko* company (Afonso et al., 2012). This software is also used for the data analysis using methods of PCA extended to symbolic data. This extended PCA offers many features for visualization as the projection of histograms and interval values in the factorial plane as well as the projection of clusters of countries performed with k-means clustering also extended to symbolic data (see Figure 1).

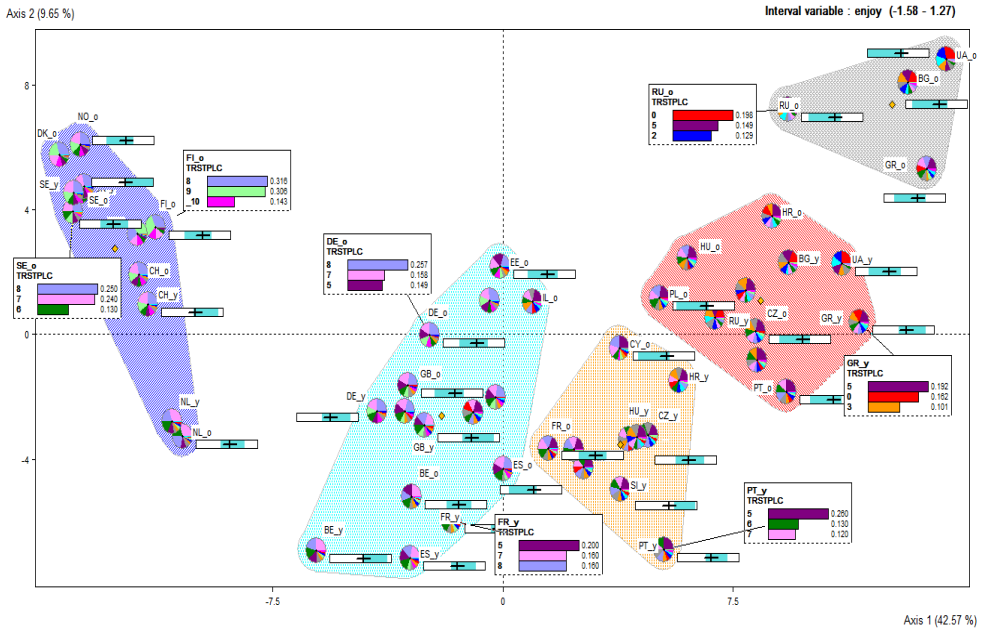


Figure 1: PCA factorial plane of the higher level individuals "country x age" with SYR software. Symbolic variables and Clusters of higher level individuals can be visualized in the factorial plane.

The example of Figure 1 uses the micro data of the round 5 of the European Social Survey (ESS). The aggregates are the cross-classifications by 26 countries and two age groups so that age ='y' corresponds to 'less than 50 years old,' and age = 'o' corresponds to '50 years old at least,' respectively. There are thus 52 symbolic objects in the data. The symbolic variables, respectively, are of the following two types: (i) attitudes, opinions etc of survey questions with 10 categories, or (ii) intervals on people's life values so that the interval range is from 25% quartile to 75% quartile.

In the Figure, we visualise 5 clusters of individuals "country x age". These clusters were obtained with k-means extended to symbolic data (ClustSyr) applied to the coordinates of the points in the factorial plane. Scandinavian countries are all at the upper left of the factorial plane with Switzerland. Netherland is also in the same cluster but at the lower left of the factorial plane. There is a cluster at the upper right with only people over 50 years old from Russia (RU), Ukraine (UA), Bulgaria (BG) and Greece (GR). Western countries (except Portugal and Greece) are at the middle of the plane. Portugal is at the right with eastern countries. Younger people and older people from France or Portugal are not in the same clusters. In the picture, we also visualize the histogram variable TRSTPLC (trust in Police) in the factorial plane. For each concept "country x age", we visualize this variable thanks to pie charts. By clicking on each pie chart, we obtain the details of each histogram. We note the very bad results of Russia, Bulgaria and Ukraine at the upper right (with scores equals to 0/10, 2/10, 5/10) and the very good results of Finland, Switzerland, Denmark, Sweden, Norway at the left (with scores equals to 8/10, 9/10, 10/10). In this factorial plane, we finally visualize the interval-valued variable « enjoy ». Each value is indicated by a blue rectangle in full within an empty black rectangle. The empty black rectangle indicates the min of the min and the max of the max among all the concepts. We also note that the results are getting better from the upper right to the upper left but this trend is not as clear as for the variable "TRSTPLC".

References

Afonso, F., Haddad, R., Toque, C., Eliezer E.-S., Diday, E. (2012). *User Manual of the SYR Software,* Syrokko internal publication, 70pp., http:// www.syrokko.com

Laaksonen, Seppo (2008). People's Life Values and Trust Components in Europe - Symbolic Data Analysis for 20-22 Countries. In. *Edwin Diday and Monique Noirhomme-Fraiture, "Symbolic Data Analysis and the SODAS Software", Chapter 22*, pp. 405-419. Wiley and Sons: Chichester, UK.

Laaksonen, Seppo (2010). The Survey as a Basis for Symbolic Data Analysis. In: *Official Statistics , Methodology and Applications in Honour of Daniel Thorburn* (Eds. Michael Carlson, Hans Nyquist and Mattias Villani), pp. 93-106.

The ESS data archive: *Online; accessed 5 -Sept-2012]*. http://ess.nsd.uib.no/ess/round5/

Interval and classic time series forecasting combination system. Applications to exchange rate (FOREX) prediction

Carlos Maté[1] & Laura Morell

1. ETS de Ingeniería (ICAI). Universidad Pontificia Comillas. Alberto Aguilera, 25. Madrid 28015 (SPAIN).
*Contact author: cmate@upcomillas.es

Keywords: ARIMA, combined forecast, hybrid methodology, interval-valued data, k-NN.

Having accurate forecasts of real world variables in economics, health and so on is a critical issue in the new world arising from the global crisis where resources are strongly limited but needs of people are increasing. All practitioners and professors agree that combining forecasts reduces the final forecasting error (see Bates and Granger (1969), Clemen (1989) and Timmerman (2006)). Recently, tourism, forestry or hydrology has discovered the advantages of forecasts combination when the main goal is to obtain smaller forecasting errors. One advantage of combining over not combining is that when there is a structural gap in the available information of the data to be analyzed, simple methods differ in their adaptability to the rupture of the data in terms of greater errors than those obtained in situations without structural breaks. Hence combining forecasts will be more efficient than using single forecasts.

Following Maté (2011b), we will be interested in forecasting a magnitude, denoted by *Y*, for a large number of forecast horizons, denoted by *n*, from the information about it in different periods of time 1, 2, … until the present moment (denoted by *t*). With this objective in mind, the forecaster has several forecasts obtained using a number of forecasting methods, denoted by *p*. Hence, we will assume that we have *p* unbiased forecasts of the magnitude *Y* obtained by *p* forecasting methods (or experts or a combination of methods and experts) $\left(M_1, M_2, ..., M_p\right)$ for *n* forecast horizons $\left(H_1, H_2, ..., H_n\right)$. The notation $f_{ij}(t)$ stands for the forecast obtained in the ith forecast horizon by the jth method or expert with $i=1,...,n$ and $j=1,...,p$. If we have the above *p* unbiased forecasts of the same variable or magnitude Y for the same time horizon (let this be *i*), then the composite forecast, denoted by $c_i(t)$, based upon the $p\times 1$ vector of the linear weights, $w_i(t)$, will adopt the expression $c_i(t)=w_i^T(t)f_i(t)$ where $\sum_{j=1}^{p} w_{ij}(t)=1$. A very important and particular case is when $w_{i1}(t)=...=w_{ip}(t)=\frac{1}{p}$, providing as combined forecast the simple arithmetic mean of the p forecasts. This combination is the one most widely used, and for some authors the best, although some other authors consider the minimum variance method as more accurate.

The use of interval analysis (IA) represents a new research line in forecasting in the last 7 years (see Maté (2011a) for a brief history of IA). The introduction of interval time series (ITS) concepts and forecasting methods has been proposed in different papers (Arroyo et al. (2011), Arroyo and Maté (2009), García-Ascanio and Maté (2010) and Maia et al. (2008), among others).

Given that we are able to use several forecasting methods with ITS, one main issue in SDA with interval-valued data is how to combine several forecasts obtained for an ITS. This paper will show the new tool ITSFOCOMB (Interval Time Series Forecasting Combination System) to obtain forecasts of

ITS. It allows the user to combine forecasts as well as to forecast using key forecasting strategies based on models like ARIMA and non-based on models like kNN. Several cases have been run using ITS obtained from FOREX. In particular, two exchange rates have been considered:

- Case A. The classic daily series USDJPY (dollar-yen), from 01/04/1971 to 12/20/2010;
- Case B. The classic daily series EURUSD (euro-dollar), from 01/01/1999 to 12/20/2010.

Every time series is converted to a monthly ITS (with ITSFOCOMB) in order to obtain the forecasts for the months January 2009 to December 2010 (case A) and January 2010 to December 2010 (case B). In these cases we analyze the forecasters choice between forecasting directly with interval data or forecasting with the classic series and then to transform these forecasts into intervals. A comparison between different classic forecasting methods and a comparison between different interval forecasting methods included in the tool has been made, in order to clarify which method is the best one and gives the smallest forecasting errors. The main conclusion is that the best option is to forecast using ITS and interval forecasting methods. Other set of cases studies the hybrid forecasting methodology, forecasting the series of the centers using the best forecasting model available (ARIMA) and then separately, the series of the radius with the best forecasting method (kNN) to then form the interval forecasted series and study the errors. In general, hybrid methods give better results than the other forecasting methods. This result confirms the findings in Zhang (2003) and in Maia et al. (2008).

Given that both combining forecasts and hybrid methodology have separately obtained promising results with FOREX data in an ITS context, one issue to be answered is the following. What is, in terms of error measures, the effect of combining forecasts for ITS obtained by hybrid methodology over not combining? We will give some answers with the above series.

References

Arroyo, J., Maté, C. (2009). Forecasting histogram time series with k-nearest neighbours methods. *International Journal of Forecasting*, 25 (1), 192-207.

Arroyo, J., Espínola, R., Maté, C. (2011). Different approaches to forecast interval time series: a comparison in Finance. *Computational Economics*, 37 (2), 169-191.

Bates, J.; Granger C. (1969). The combination of forecasts. *Operations Research Quarterly*, **20**, 451-468.

Clemen R.T. (1989). Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting*, **5**, 559-584.

García-Ascanio, C.; Maté, C. (2010). Electric power demand forecasting using interval time series: A comparison between VAR and iMLP. *Energy Policy* 38, 715-725.

Maia, A.L.S., De Carvalho, F.D.A.T., Ludermir, T.B. (2008). Forecasting models for interval-valued time series. *Neurocomputing*, 71 (16-18), 3344-3352.

Maté, C. G. (2011a). Book Review: Ramon E. Moore, Ralph B. Kearfott, Michael J. Cloud (Eds.), Introduction to interval analysis, SIAM, USA, 2009, 223pp., ISBN: 978-0-898716-69-6 (hardcover), $72. *Fuzzy Sets and Systems*, Vol. 177, Issue 1, 95-97.

Maté, C.G. (2011b). A multivariate analysis approach to forecasts combination. Application to foreign exchange (FX) markets [Una aproximación a la combinación de pronósticos basada en técnicas de análisis multivariante]. *Revista Colombiana de Estadistica*, 34 (SPEC. ISSUE 2), 347-375.

Timmermann A. (2006). Forecast combinations. In *Handbook of Economic Forecasting*, 135-196. Elliott G., Granger C. W. J. and Timmermann A. (eds); Elsevier: Amsterdam.

Zhang, G.P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, 159-175.

Session IX

Clustering II

November 9, 11:30 - 13:00

A hierarchical clustering algorithm applied to teacher ratings described by modal ordinal symbolic data

Carmen Bravo[1,*], José M. García-Santesmases[2]

1. Universidad Complutense de Madrid, Servicio Informático de Apoyo al Usuario-Investigación, Edificio Vicerrectorado de Alumnos, 28040 Madrid, Spain.
2. Universidad Complutense de Madrid, Facultad de Ciencias Matemáticas, Dpto. Estadística e Investigación Operativa, 28040 Madrid, Spain.
*Contact author: mcbravo@pas.ucm.es

Keywords: Modal ordinal symbolic data, symbolic consensus measure, symbolic hierarchical clustering.

An ascending hierarchical clustering algorithm (Ward, 1963) is applied to objects described by modal ordinal symbolic data (Bock and Diday, 2000). The criterion to be minimized in each step is based on the decrease of a variability measure of one partition when two of its members are joined. This decreasing value is based on a consensus measure which is proportional to the dissimilarity between the two clusters joined.

We establish a general φ function that characterizes a consensus measure (Tastle et al., 2005) defined for probability distributions for a set of ordinal categories. We extend this measure to sets of modal ordinal symbolic data objects and define a dissimilarity measure between two of these sets based in the consensus variability of their centroids. Previous work regarding consensus measures for modal ordinal symbolic data is in García-Santesmases and Bravo (2010) and in García-Santesmases et al. (2010). In the present work we use the Leik measure (Leik, 1966) as the φ function.

To illustrate the proposed method we apply it to a data set composed of 34 teachers described by modal ordinal symbolic data. This new approach allows whatever ordinal scale. Teachers were rated by their students (1350) on 12 items on a 1 to 4 scale meaning: 1 poor, 2 average, 3 good, 4 excellent. The observed items are: initial subject presentation; teacher setting to course syllabus; well time management; evoking interest in the students about the subject; use of practical examples; stimulating students to be active in class and readiness to clear their doubts; readiness to give advice in academic development; degree of respect between students and teacher; subject knowledge; stimulating students to read books, journals and magazines; communications skills; and, ability to clear students' doubts.

Some criteria to measure the quality of partitions and clusters are given. Interpretations of clusters regarding relevant issues are also shown.

References

Bock, H.H., Diday, E. (Eds.) (2000). *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer Verlag, Heidelberg.

Garcia-Santesmases, J.M., Bravo, M.C. (2010). Consensus Analysis through Modal Symbolic Objects, In: *Compstat 2010 proceedings*, Springer, ISBN 978-3-7908-2603-6, 1055--1062.

García-Santesmases, J.M., Franco C., Montero J. (2010). Consensus Measures for Symbolic Data. *Computer Engineering and Information Science*, 4, 651--658.

Leik, K.R. (1966): A measure of ordinal consensus. *The Pacific Sociological Review*, 9, 85--90.

Tastle, W.J., Wierman, M.J., Dumdum, U.R. (2005). Ranking Ordinal Scales Using the Consensus Measure. *Issues in Information Systems*, 6(2), 96--102.

Ward, J. (1963), Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236--244.

Clustering based on normal mixture model for aggregated symbolic data

Nobuo Shimizu[1,*] , Junji Nakano[1]

1. The Institute of Statistical Mathematics, Japan
*Contact author: nobuo@ism.ac.jp

Keywords: Cluster analysis, EM algorithm, Normal mixture model, Symbolic data

For clustering symbolic data (SD), hierarchical methods based on several definitions of dissimilarity between different two SD have been studied in symbolic data analysis (SDA) (Billard and Diday, 2006). As mixture model-based clustering methods for classical data are becoming popular recently (Everitt et al., 2011), we investigate clustering method based on normal mixture model in SD framework.

Traditional SDA uses information only about marginal distribution of each variable in each SD. We consider the case where individuals of classical data are divided into some natural defined groups and each group is considered to be SD, and call it aggregated symbolic data (ASD). ASD can be represented by information about its marginal distributions and also by information about joint distribution.

EM algorithm (Dempster et al., 1977) is often used in model-based clustering for classical data, and various extensions to some mixture models of EM algorithm are studied (McLachlan and Peel, 2000 ; McLachlan and Krishnan, 2008). We derive simplified EM algorithm in clustering based on normal mixture model for ASD by using mean and variance of each variable in ASD and also covariance information among variables in ASD. We apply our method to artificial and real data examples.

References

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining.* John Wiley & Sons Ltd, Chichester, UK.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B 39*, 1–38.

Everitt, B. S., Landau, S., Leese, M. and Stahl, D. (2011). *Cluster Analysis, 5th Edition.* John Wiley & Sons Ltd, Chichester, UK.

McLachlan, G. J. and Krishnan, T. (2008). *The EM algorithm and Extensions, Second Edition.* John Wiley & Sons Inc, Hoboken, NJ.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models.* John Wiley & Sons Inc, New York.

Clustering II

Some partitioning fuzzy clustering algorithms for interval-valued data

Francisco de A. T. de Carvalho[1,*]

1. Centro de Informatica - CIn/UFPE
*Contact author: fatc@cin.ufpe.bf

Keywords: Fuzzy clustering, Interval data, City-Block distances, Hausdorff distances, Symbolic data analysis

This presentation aims at giving partitioning fuzzy clustering algorithms in order to cluster objects described by interval-valued variables. Interval-valued variables are needed, for example, when an object represents a group of individuals and the variables used to describe it need to assume a value which express the variability inherent to the description of a group. Interval-valued data arise in practical situations such as recording monthly interval temperatures at meteorological stations, daily interval stock prices, etc. Another source of interval-valued data is the aggregation of huge databases into a reduced number of groups, the properties of which are described by interval-valued variables. Therefore, tools for interval-valued data analysis are very much required (Bock and Diday, 2000) .

Symbolic data analysis has provided suitable tools for clustering symbolic data: agglomerative and divisive hierarchical methods, partitioning hard cluster algorithms. Concerning fuzzy clustering of interval-valued data, El-Sonbaty and Ismail (1998) presented a fuzzy K-means algorithm for clustering data on the basis of different types of symbolic variables. Yang et al (2004) presented fuzzy clustering algorithms for mixed features of symbolic and fuzzy data. In these fuzzy clustering algorithms, the degree of membership is associated to the values of the features in the clusters for the cluster centers rather than being associated to the patterns in each cluster, as is the standard approaches. De Carvalho (2007) presented fuzzy clustering algorithms based on Euclidean distances and De Carvalho and Tenorio (2010) introduced fuzzy clustering algorithms based on adaptive quadratic distances.

Chavent and Lechevallier (2002) introduced a hard clustering algorithm based on non-adaptive Haudorff distances between vectors of intervals. De Souza and De Carvalho (2004) gives adaptive and non-adaptive hard clusterings algorithms based on City-Block distances. Later, De Carvalho et al. (2006) and De Carvalho and Lechevallier (2009) proposed an adaptive hard clustering algorithm based on Hausodorff and City-Block distances for interval-valued data.

This paper extends these works to give partitioning fuzzy clustering algorithms for interval-valued data based on adaptive and non-adaptive City-Block and Hausdorff distances. Conventional hard clustering methods restrict each item of the data set to exactly one cluster. Fuzzy clustering generates a fuzzy partition based on the idea of partial membership expressed by the degree of membership of each item in a given cluster. Thus, fuzzy clustering techniques allow the user to distinguish between objects which are strongly associated with particular clusters from those that have only a marginal association with multiple clusters (Bezdek, 1981).

These partitioning fuzzy clustering algorithms are related to the dynamical clustering algorithm and are iterative two (representation and allocation) or three (representation, weighting and allocation) steps relocation algorithms that looks for a partition of a set of objects into a fixed number of clusters and their corresponding prototypes such that a clustering criterion (objective function) measuring the fit between the clusters and their representatives is locally minimized. These steps

are repeated until a clustering stopping criterion is reachead. In this presentation, for each method, it is given the clustering criterion (objective function) and the main steps of the algorithms (the computation of the best prototypes in the representation step, the computation of the best relevance weights of the variables if there is a weighting step, and the determination of the best partition in the allocation step). The performance, robustness and usefulness of these fuzzy clustering algorithms are illustrated with real interval-valued data sets.

References

M. Chavent and Y. Lechevallier (2002). Dynamical clustering algorithm of interval data: optimization of an adequacy criterion based on Hausdorff distance. In *IFCS 2002, 8^{th} Conference of the International Federation of Classification Societies (Cracow, Poland)*, pp. 53–59.

J. C. Bezdek (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms,* Plenum Press, New York.

H-.H. Bock and E. Diday (2000). *Analysis of Symbolic Data,* Springer, Berlin et al.

F.A.T. De Carvalho (2007). Fuzzy c-means clustering methods for symbolic interval data. *Pattern Recognition Letters, 28*, 423–437.

F.A.T. De Carvalho and C.P. Tenorio (2010). Fuzzy K-means clustering algorithms for interval-valued data based on adaptive quadratic distances. *Fuzzy Sets and Systems, 141*, 301–317.

F.A.T. De Carvalho, R.M.C.R. Souza, M. Chavent, and Y. Lechevallier (2006). Adaptive Hausdorff distances and dynamic clustering of symbolic data. *Pattern Recognition Letters, 161*, 2978–2999.

F.A.T. De Carvalho and Y. Lechevallier (2010). Partitional clustering algorithms for symbolic interval data based on single adaptive distances. *Pattern Recognition, 42*, 1223–1236.

Y. El-Sonbaty and M.A. Ismail (1998). Fuzzy clustering for symbolic data. *IEEE Transactions on Fuzzy Systems, 6*, 195–204.

R.M.C.R. De Souza and F.A.T. De Carvalho (2004). Clustering of interval data based on City-Block distances. *Pattern Recognition Letters, 25*, 353–365.

M.-S. Yang, P.-Y. Hwang, and D.-H. Chen (2004). Fuzzy clustering algorithms for mixed feature variables. *Fuzzy Sets and Systems, 141*, 301–317.

Authors

Afonso, Filipe (France)
Amouh, Teh (Belgium)
André, Maia (Brazil)
Araújo, Júnior (Carlos Brazil)
Arroyo, Javier (Spain)
Batagelj, Vladimir (Slovenia)
Billard, Lynne (United States of America)
Bravo, Carmen (Spain)
Brito, Paula (Portugal)
Chavent, Marie (France)
Chen, Chun-houh (Taiwan)
de Barros, Alberto (Brazil)
de Carvalho, Francisco (Brazil)
Dias, Sónia (Portugal)
Diday, Edwin (France)
Drago, Carlo (Italy)
Duarte Silva, Pedro (Portugal)
García-Santesmases, José (Spain)
Giordano, Giuseppe (Italy)
Groenen, Patrick (Netherlands)
Guijarro, María (Spain)
Haddad, Raja (France)
Irpino, Antonio (Italy)
Kao, Chiun-How (Taiwan)
Kejžar, Nataša (Slovenia)
Korenjak Černe, Simona (Slovenia)
Laaksonen, Seppo (France)
Lauro, Carlo (Italy)
Lima Neto, Eufrasio (Brazil)
Maté, Carlos (Spain)
Minami, Hiroyuki (Japan)
Mizuta, Masahiro (Japan)
Morell, Laura (Spain)
Murillo, Juan de Dios (Costa Rica)
Nakano, Junji (Japan)
Noirhomme-Fraiture, Monique (Belgium)